JOURNAL OF SOCIAL COMPUTING
ISSN 2097-5015 02/07 pp287–299
Volume 6, Number 4, December 2025
DOI: 10.23919/JSC.2025.0024

Structural Synchronization in Classification of Controversial Content

Douglas Guilbeault* and Damon Centola*

Abstract: Morally controversial content, such as offensive and hateful images over social media, is especially challenging to categorize, given widespread disagreement in how people interpret and evaluate this content. Numerous studies argue that a range of subjective biases, such as partisan differences in moral reasoning, lead people not only to diverge in their classifications of controversial content, but also to resist any attempts to change their classification judgments via social influence. Yet, recent large-scale analyses of classification patterns over social media suggest that separate populations, such as democrats and republicans, can reach surprising levels of agreement in the categorization of inflammatory content like fake news and hate speech, despite considerable differences in their moral reasoning and worldview. This poses a fundamental puzzle: how can populations of diverse individuals who disagree in the interpretation of controversial content nevertheless arrive at highly similar decisions for the classification and removal of such content? Here, we use an online platform to test the hypothesis that structural symmetries in information exchange networks can synchronize convergence on decisions regarding the classification and removal of controversial images across independent networks, leading them to independently reproduce consistent systems of classification. We find that isolated individuals diverge considerably in their classification of controversial content, whereas separate, structurally similar networks independently synchronize in their classifications and content removal decisions, reducing partisan biases across all networks. We also find that when participant experience is compared to subjects evaluating content individually in the control condition, participants within synchronizing networks reported having significantly more positive feelings about their task, and experience significantly less emotional stress when evaluating controversial content.

Key words: classification; coordination; social networks; online experiment; content moderation; collective intelligence; crowdsourcing

1 Introduction

Individuals vary substantially in how they categorize novel phenomena[1–3] and particularly in how they interpret emotionally charged[4, 5] and morally controversial content[6] such as offensive and hateful images over social media. This individual variation is captured not only by experimental research[2, 3, 7], but also by qualitative and quantitative analyses of people who work as content moderators specializing in the practice of categorizing and removing controversial content that violates the community standards of social media[8–14]. Despite the urgency of content moderation, which has been recognized by both

• Douglas Guilbeault is with Graduate School of Business, Stanford University, Palo Alto, CA 94305, USA. E-mail: dguilb@stanford.edu.
• Damon Centola is with Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA 19104, USA. E-mail: dcentola@asc.upenn.edu.
* To whom correspondence should be addressed.
Manuscript received: 2025-07-26; revised: 2025-10-30; accepted: 2025-11-10

Congress[15] and the United Nations[16], research shows that even trained content moderators consistently disagree in their decisions for how to classify hate speech and offensive images, and whether to remove this content[8, 10–14, 17, 18]. Yet, puzzlingly, recent large-scale analyses of classification patterns over social media find that distinct populations of internet users (e.g., Democrats and Republicans) can arrive at similar categorizations of inflammatory content, such as partisan news and hate speech[19–26], despite considerable differences in their political orientations and moral reasoning[27–29]. This raises an important puzzle: How can separate and diverse populations of social media users arrive at similar ways of categorizing controversial content, despite considerable differences in their interpretation of and approach to such content?

Here, we provide novel insights into this puzzle by investigating how interacting in social networks can filter individual variation and lead separate networks (or "teams") to arrive at highly similar classifications of controversial content, a phenomenon we refer to as structural synchronization. Our theory is motivated by recent studies of collective intelligence which find that exchanging information in structured social networks reduces variation in individual judgments within populations, while also leading to convergent solutions across populations[30–37]. In particular, a recent study by Guilbeault et al.[7] in 2021 finds that distinct populations with similar network structures independently evolve highly similar classification schemes for categorizing novel visual stimuli—viz., the same stimuli for which isolated individuals produced highly varied and inconsistent classification schemes. This prior work emphasizes the role of structure—and specifically, the structural feature of scale—as the driving mechanism behind synchronization, since these prior studies find that interactions in small groups are insufficient for inducing replicable patterns of classification. Guilbeault et al.[7] showed that while individuals share preferences for particular classifications—due to shared psychological processes or shared culture—these preferences are too weak to enable synchronization at small scales. They find that synchronization only emerges in large social networks, since these large-scale structures enable weakly preferred categories to achieve sufficient critical mass to outcompete alternative classifications, leading these categories to consistently spread and gain adoption. These findings suggest a general hypothesis about the influence of population structure—and in particular, scale—on belief synchronization across independent populations.

Unfortunately, prior work in this area focused on arbitrary content without any political or social implications. Thus, it remains unknown whether network structure may also engender social synchrony on vital topics relating to the classification and interpretation of controversial and potentially harmful content online (e.g., posts on social media). It is frequently argued that morally offensive and politically charged content activates motivated reasoning which leads people to be highly resistant to changing their views, suggesting that the benefits of structural mechanisms may break down in such cases[38–45]. The key mechanism underlying this view is the idea that exposure to conflicting information activates motivated reasoning[42, 46, 47], for example, by triggering people to develop and defend arguments that protect their existing views and avoid cognitive dissonance caused by holding contradictory beliefs[48, 49].

However, more recent work suggests that motivated reasoning may be primarily driven by the activation of social identity related information[40–45, 50] rather than exposure to contradictory views themselves—for example, by triggering people to form arguments that maintain their membership within a particular group through a conformity strategy. The implication of this view is that, in the absence of salient intergroup boundaries and dynamics, individuals may be significantly more receptive to countervailing information and willing to update their beliefs accordingly. Indeed, recent work on collective intelligence suggests that social influence in structured social networks can lead individuals to overcome psychological biases and converge on accurate and replicable judgments[7, 30, 32, 51]. This includes research showing that, in the absence of partisan identity information, both Democrats and Republicans can learn from each other and converge on similar interpretations of data in politically-charged forecasting tasks[30, 32, 51]. Yet, no prior work to our knowledge has extended these findings to the domain of controversial content classification; it is not clear that

the same convergence dynamics will occur in the interpretation and classification of salient and controversial images, since images are expected to have a particularly strong effect in activating psychological biases[52, 53]. Building on this work in collective intelligence, we hypothesize that structured social networks will promote synchrony (i.e., convergence) across independent populations, leading them to reach similar decisions about the classification and removal of controversial material[7, 30, 32].

We use a pre-registered experiment to test our hypothesis (https://osf.io/v6dme). In every replication of our study, we compare classifications of independent individuals (i.e., control condition) with the classification judgments of networked participants, all evaluating the same content. We hypothesize that independent individuals will exhibit diversity both in how they classify controversial content, and in whether they decide to remove this content. By contrast, we predict that collaborative classification efforts in structured communication networks will produce emergent synchronization across all of the networks, leading every network to exhibit greater consistency and quality in their judgments than participants in the control condition.

2 Material and Method

Experimental design. We recruited 620 subjects from Amazon's Mechanical Turk (Mturk) to participate in a paid online content moderation task called the "Facebook Flagging Task". This research was funded by Facebook's Content Moderation Research Award. Facebook played no role in the design of this study, nor in the collection and analysis of the data. The sample of participants in our study was drawn from individuals who work in content moderation across a range of public social media platforms. Across all experimental conditions, participants were instructed that the images in their task were drawn from the posts of Facebook users, and that Facebook was requesting participants' assistance in determining whether or not the displayed images should be removed from their platform. All participants were adult US citizens and active social media users who provided informed consent, with full knowledge that the task would involve evaluating the appropriateness of potentially upsetting images. There were no differences in subject demographics across conditions.

Subjects were randomized into one of two conditions: (1) the "individual" condition, in which independent individuals classified and removed images, without any communication with other content moderators; and (2) the "network" condition involved socially interactive classification, in which fifty individuals collaboratively classified and removed images by coordinating and sharing information in structured peer networks. 220 participants were randomized to the "individual" condition; 400 participants were randomized to the "network" condition, forming 8 independent social networks, with no overlap in the sample of subjects between networks. All analyses to follow are at the trial level to control for statistical nonindependence among content moderators in networked teams. See the "Material and Method" section in the Electronic Supplementary Material (ESM) in the online version of this article for more extensive description of the methodological design and analytic approach.

Subject experience. Subjects were tasked with classifying a large set of controversial images collected from social media sites (see Figs. S1 and S2 in the ESM). Subjects were exposed to a set of images randomly drawn from a continuum of 600 images, ranging from depictions of interpersonal conflict (including domestic violence and bullying) to depictions of militaristic violence (including armed conflict and terrorism). Subjects were exposed to a random selection of images from this continuum over the course of the study. There were no significant differences in the set of images seen by subjects across experimental conditions.

Participants were instructed that if they deemed an image to be inappropriate—such that it should be removed from Facebook—they should indicate this by selecting a "violation tag" for the image from a drop-down menu of 16 options (see Fig. S1 in the ESM). The violation tags were drawn from Facebook's official Guidelines and Community Standards[8]. Subjects were also told that if they believed that an image should not be removed, they should select the "Do Not Remove" option from the drop-down menu. The drop-down menu options were identical across experimental conditions. The order of the options in the drop-down menu was randomized in each round and for each subject in both experimental conditions to prevent

order effects (see Fig. S1 in the ESM).

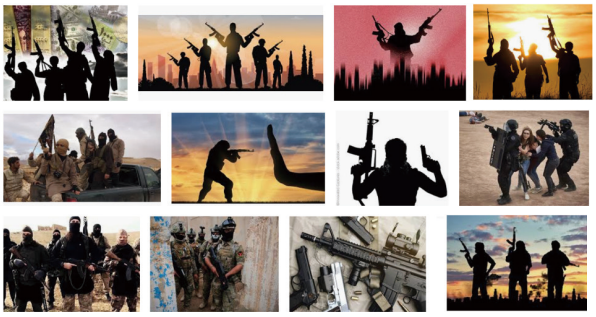
In the individual condition of the Facebook Flagging Task, subjects adopted a common approach to content moderation, currently used by Facebook and many leading social media platforms[8, 9, 18]. Individuals viewed images one at a time without knowing the source or context of the image (see Fig. S1 in the ESM)[8, 9, 18]. Participants were given identical instructions in the networked version of the task, except in this condition subjects were asked to coordinate with their network neighbors to collaboratively decide how to classify content and whether to remove it. Subjects in the network condition participated in a sequence of pairwise interactions with other content moderators in their fifty-person networks. All networks were structurally identical: complete, fully-connected graphs composed of fifty unique participants. Each round in the network condition of the Facebook Flagging task proceeded as follows[7].

Each round began with participants being randomly paired with a member of their network (resulting in twenty-five pairings per round in each network). In each pair, one participant was randomly assigned to be the "speaker" and the other the "hearer". The speaker in each pair was shown a set of three randomly selected images from the continuum (see Fig. S2 in the ESM), in which one of the three images was highlighted. Then, the speaker was given thirty seconds to make a selection from the drop-down menu that would enable their partner to distinguish the highlighted image from the other two presented images. The speaker could either select a violation tag from the list Facebook content violation tags (for instance, sexism, racism, etc.), or, if the speaker deemed the image not to be a violation, they could select the "Do Not Remove" option. Finally, the hearer was then shown the same set of images in a randomized order (without any images highlighted), along with the violation tag that their partner had selected. The hearer was then given thirty seconds to identify the image corresponding to the speaker's assigned tag. If the hearer succeeded in selecting the correct image, both participants were compensated with a cash payment and both were informed about their agreement. However, if the hearer failed to select the correct image, both participants lost money and both participants were shown their partners' selections. Once all pairs completed a round, a new round would begin with everyone in the network being randomly paired again.

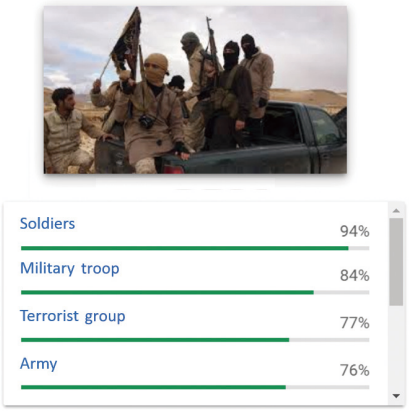
Participants received no information about the decisions of other members of the population; they only had access to their partner's response in the round in which they were paired[28, 34]. All interactions were pairwise and anonymous. Subjects did not have information about their partner's identity, nor the size of their network. The instructions for participants, and the stimuli to which they were exposed, were identical across experimental conditions. Consequently, any differences across experimental conditions in the observed patterns of content classification and removal can be attributed to the direct effects of networks on content moderation.

Subject recruitment. All participants were recruited from Amazon's Mechanical Turk. In order to participate, recruits were required to be US citizens and active users of social media with English as their first language. Subjects were also asked whether they possessed prior experience working as a content moderator for a social media platform, and only those subjects with prior experience as a content moderator were invited to participate in this task. 55.3% identified as male and 44.7% identified as female. 49.6% identified as democrat, 28.3% as independent, 20.7% as republican, and the remaining identified as belonging to an "Other Party". 75.5% identified as white, 9.9% as Black, 8.4% as Asian, 4.2% as Mixed, 1.1% as Other, less than 1% as American Indian or Alaska Native, and less than 1% as Native Hawaiian or Pacific Islander. In terms of ethnicity, 90.6% selected "none", 5.3% selected Hispanic, 2.7% selected Latino, and 1.2% selected Spanish. There were no significant differences in the distribution of demographic traits across conditions. All participants were legal adults and provided written consent with full knowledge that they would encounter upsetting images as part of the content moderation task. Data were collected between August 2019 and February 2020.

The methodological steps are as follows. Step 1: Using compsyn to scrape images from Google with the search terms "abuse" and "terrorism" (see Fig. 1a); Step 2: Using Google Vision to classify images in terms of their degree of membership in the respective categories, "abuse" and "terrorism" (see Fig. 1b); Step



(a) Scrape images from Google



(b) Classify images using Google Vision

(c) Arrange images into a semantic continuum

Fig. 1 Schematic display of our methodology.

3: Using Google Vision classification rankings to arrange images into a linear semantic continuum (see Fig. 1c), such that moving leftward along the continuum increases the likelihood that images are recognized as depicting "abuse", and rightward movement along the continuum increases the likelihood that images are recognized as depicting "terrorism". In the middle of the continuum are ambiguous images that are equally and weakly associated with abuse and terrorism.

Constructing semantic continuum. The images were selected and arranged using a hybrid approach, involving machine learning and human crowdsourcing (Fig. 1a). Following recent work[54], this method was designed to construct a semantic continuum using images that allowed us to identify how individuals and networks grouped together the images using violation tags, and also to identify the accuracy of the violation tags that groups used. First, we scraped the top 500 images from Google based on two focal concepts that are frequent violations of Facebook's community standards, namely: abuse and terrorism (Fig. 1a). Second, we classified the extent to which each image belonged to each focal category using Google Vision, a machine learning ensemble that applies feature detection and online metadata to label images and score their membership in each label (Fig. 1b). We used Google Vision's classifications to arrange images into a linear semantic continuum, so that images in the leftmost pole of the continuum were increasingly associated with abuse, while images in the rightmost pole were increasingly associated with terrorism (Fig. 1c). The middle of the continuum contained ambiguous images that were less strongly associated with abuse or terrorism, but which contained visual features relating to both, such as images depicting police brutality or political protests. Low quality or irrelevant images were removed, along with images that included branding, resulting in a continuum of 600 images. Images in the 1–200 range were associated with abuse; images in the 200–400 were ambiguous and weakly related to abuse and terrorism; and images in the 401–600 ranges were associated with terrorism.

As a robustness check, we used human crowdsourcing to verify that our arrangement of the images successfully created a linear semantic continuum (Fig. 2). The data from the individual condition were used to demonstrate that subjects are more likely to choose the label abuse for images further to the left of the continuum, and subjects are more likely to choose the label terrorism for image further to the right of the continuum (Fig. 2a). As images approach the middle of the continuum, individuals are less likely to use either abuse or terrorism, and overall variation increases among individuals in terms of the labels they chose (Fig. 2b). Importantly, even though the image continuum is structured to facilitate the identification of boundaries in grouping together content—and even though individuals are selecting options from a finite set of predetermined choices—we observe a substantial amount of variation among individuals in terms of the labels they ascribe to

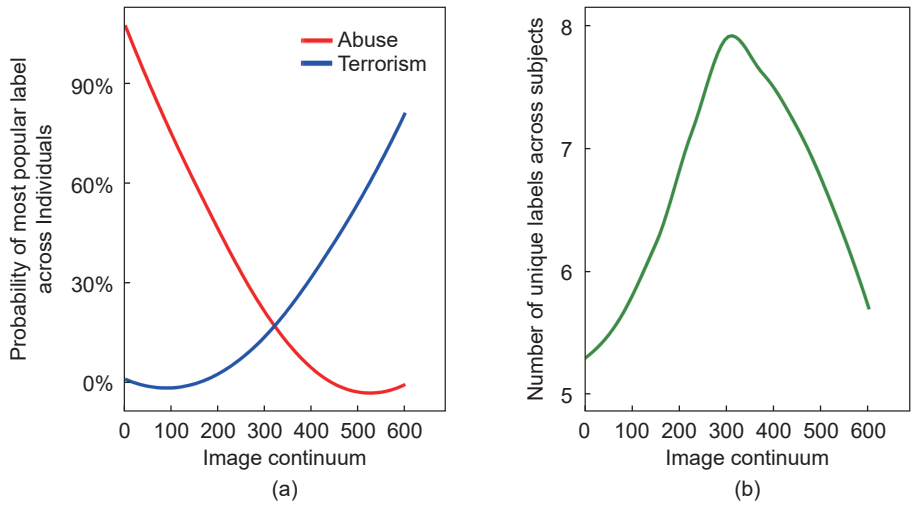


Fig. 2 Design validation of the image continuum using crowdsourcing. (a) Probability that either label, abuse or terrorism, appeared as the most popular label across independent human coders for each image in the continuum. (b) Ambiguity of images in the continuum by plotting the number of unique labels that appeared across independent human coders for each image in the continuum.

all regions of the continuum, as illustrated by Fig. 2b.

Identifying violation tags. The semantic continuum of images is designed to encode three main regions of interesting: the abuse region, from image 0 to 200; the terrorism region, from images 400 to 600; and the ambiguous region, from 200 to 400 (see "Constructing semantic continuum"). Given the pre-existing delineation of three general regions, we compare conditions in terms of the top three labels that are most frequently and successfully used throughout the task. In the individual condition, we identify the top three labels used by each individual and visualize the images that are referred to by these top three labels. In the network condition, we identify emergent categories by first identifying the top three labels that are most successfully used within each team; then, for each team, we visualize the distribution of coordination successes for each label, such that the peaks in the distributions indicate where each network most frequently succeeds in the use of each label.

The approach outlined above is applied to identify the places in the continuum where trials in each condition are more likely to apply the "Do Not Remove" option. We predict that isolated subjects in the individual condition would vary significantly in which regions of the continuum act as the centroid[55] for their use of the "Do Not Remove" option (centroid here indicates the region of peak usage of this label). By comparison, we predict that subjects in networks would be more likely to converge on using "Do Not Remove" in region the ambiguous region of the continuum (containing images that are ambiguous between abuse and terrorism; see "Constructing semantic continuum"). Since subjects use the "Do Not Remove" option throughout the continuum (though at differential rates), we employ an additional visualization technique to adjust for high variation in concentration for where subjects in the solitary condition apply this decision (since this option is used all over the continuum by solitary individuals, the density distributions without thresholding fail to accurately display underlying variation in the areas where separate individuals' use of the "Do Not Remove" option are most concentrated); specifically, rather than display the raw distribution of images that subjects' identified with the "Do Not Remove" option, we identify the median image representing the region where this option is most frequently applied, and we visualize a standard deviation of images on either side of the centroid. This adjustment is applied in displaying the data for both conditions.

Quantifying subjects' self-reported descriptions of their task experience. Subjects provide qualitative, self-reported descriptions of their task experience by completing free text entry responses to the following survey questions, which they are presented with as soon as the experiment completed: (1) How did this task make you feel? And

(2) Do you have any feedback you would like to provide about your task experience? We employ both automated and manual techniques for examining the affective tone of subjects' self-reported answers to these questions. The results of this analysis are based on automatically classifying the emotional tone of text using the popular application Linguistic Inquiry and Word Count LIWC[56] (see Fig. S3 in the ESM). LIWC employs a dictionary approach to calculating the emotional tone of text—that is, LIWC contains a list of words that are each associated with an emotional score, from −100 (negative) to 100 (positive); LIWC gathers the emotional score for each word in a given text and takes the average of these scores to provide an overall measure of sentiment. In Section 2.1 in the ESM, we show how the results are equally robust to an alternative dictionary-approach to automated sentiment classification, as well as robust to a manual approach to sentiment classification using crowdsourced human judgments (see Fig. S3 and Table S1 in the ESM).

3 Result

To begin our analyses, we examine the degree of similarity in the image classifications developed across trials within each condition (individuals vs. networks). Second, we examine the effect of communication networks on the images that moderators deem acceptable for social media (i.e., tagged "Do Not Remove"). Third, we evaluate the effects of information-sharing networks on partisan differences in the classification of controversial content. Finally, we conclude by comparing the emotional experiences of independent moderators as compared to networked content moderators.

Figure 3 shows the content tags that participants used in each condition to flag inappropriate social media content. Independent individuals exhibite substantial variation in their classification of social media content (Fig. 3; $p<0.001$, Kruskal-Wallis). However, independent networked groups classify controversial social media content with near-perfect agreement across all 8 networks (Fig. 3; $p<0.0001$, Wilcoxon Rank Sum Test, two sided).

Figure 4 shows that teams' classification schemes effectively predict their determination of acceptable social media content. Independent individuals frequently arrive at conflicting content removal decisions, exhibiting only 38% agreement in their judgments about which content should be permitted

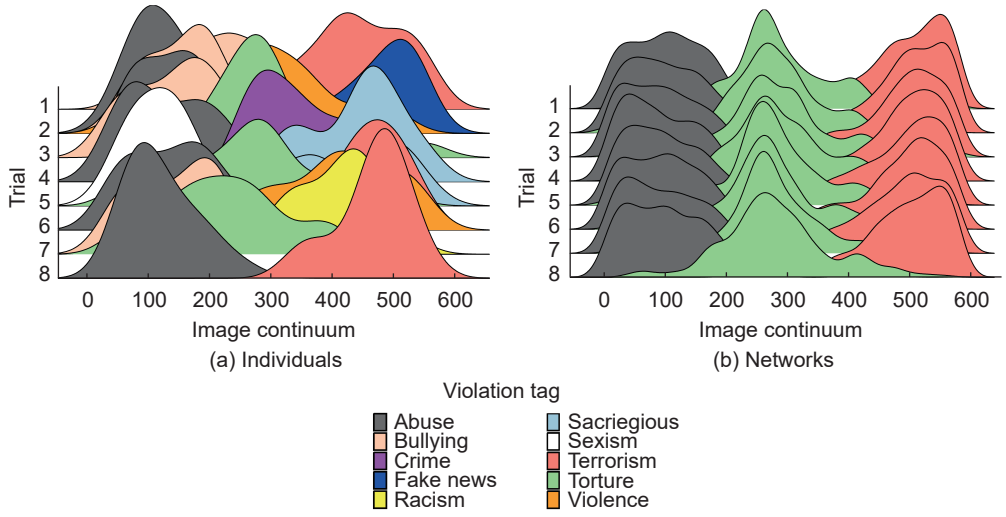


Fig. 3 A sample of the classifications that emerge in each condition. (a) Classifications that emerge among independent individuals. (b) Classification systems that emerge in separate, independent communication networks, each composed of fifty content moderators. Each row displays the classifications constructed by a single unique trial in each condition. A sample of 8 trials is shown for each condition. The horizontal axis displays the image continuum of 600 images. For trials in the individual condition, density distributions display the frequency with which specific labels are used across each region of the continuum. For network trials, the data reflect the classifications that emerge after 100 rounds of interaction, where the density distributions display the frequency of successful coordination for each label across the continuum. Each color indicates a unique label.

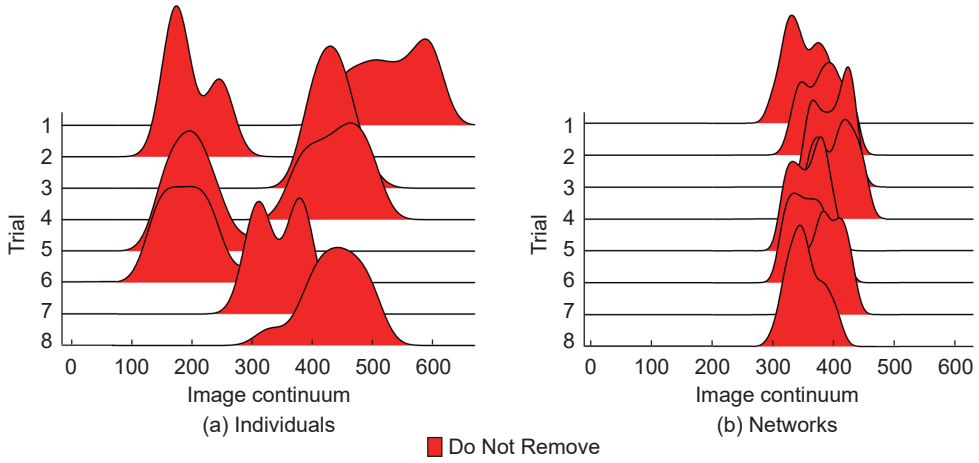


Fig. 4 Use of "Do Not Remove" option in each condition. (a) Use of this option among independent individuals. (b) Use of this option in separate, independent communication networks of fifty people. Each row displays a single unique trial in each condition. The horizontal axis indicates the full range the image continuum. For trials in the individual condition, density distributions display the frequency with which the "Do Not Remove" option is selected across each region of the continuum. For network trials, the density distributions display the frequency of successful coordination on "Do Not Remove" over 100 rounds of interaction across the image continuum.

to remain on Facebook. Strikingly, independent networked content moderation teams exhibit significantly greater consistency, reaching 64% agreement across all 8 teams in their evaluations of which social media content should be allowed to remain on Facebook (p<0.001, Wilcoxon Rank Sum Test, two-sided).

Of particular interest in these findings is their implications for partisan bias in participants' content moderation decisions. Figure 5 shows that in the independent condition on average only 30% of democrats and republicans agree in their classifications (Fig. 5, p<0.001, Proportion Test, two-sided, N=1200). By comparison, networked teams of moderators significantly reduce partisan differences in classification, leading to a 23% increase in the fraction of images for which republicans and democrats consistently agreed in their classification (p<0.001, Wilcoxon Rank Sum Test, two-sided). By the end of the task in the network condition, the majority of Democrats and Republicans agree in their image classifications. (Robustness tests reported in the ESM demonstrate that all of our main findings also hold when comparing the classification systems of networked groups to the aggregated judgments of large groups of independent moderators—often referred to as "the wisdom of the crowd"; Figs. S4–S7 and Table S2 in the ESM show that networks produce significantly more replicable and

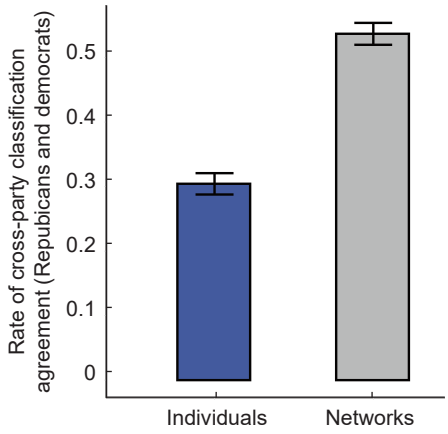


Fig. 5 Rate of cross-party classification agreement in each condition. Classification agreement indicates the fraction of images for which Republicans and Democrats agree in the most popular violation tags they assigned. Classification agreement across parties is analyzed across individuals and within networked teams to control for the non-independence of subjects in social networks. Error bars display 95% confidence intervals.

less polarized classification patterns than those produced through a common "wisdom of crowd" approach[19, 20, 57–60]).

An additional consideration for any approach to content moderation is the significant emotional stress experienced by content moderators who are exposed to large volumes of controversial and potentially disturbing social media content. To evaluate the effects

of networked versus independent content classification tasks on participants' emotional state, we administered an open-ended survey immediately concluding the study, which asked participants in each condition to provide free-text responses to two questions. The first question asked subjects to discuss how the task made them feel, and the second question asked them to provide feedback on the task. We coded each of the participants' responses using LIWC, a popular technique in natural language processes for measuring the emotional tone of text[56, 61].

Consistent with ethnographic research on content moderators[8, 9, 62, 63], responses of subjects in the individual condition indicate negative sentiment stemming from a stressful emotional experience (Fig. 6, $p<0.001$, Wilcoxon Rank Sum Test, two-sided). Yet, Fig. 6 shows that networks produce a qualitative shift in the tone of subjects' responses. In the networked teams, moderators' responses are not merely less negative, but rather significantly positive ($p<0.001$, Wilcoxon Rank Sum Test, two-sided). Moderators in the networked condition are more likely to report positive feelings relating to teamwork, productivity, and a sense of purpose for helping address an important social issue (see Table S1 in the ESM). These results are robust to a variety of automated and manual methods for coding the emotional valence of subjects' responses (see Fig. S3 and Table S1 in the ESM).

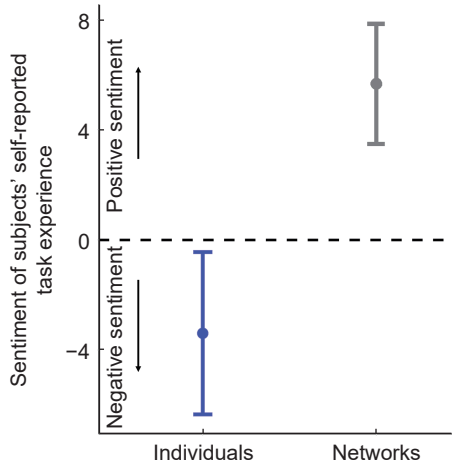


Fig. 6 Average sentiment of subjects' self-reported task experience in each condition, measured using the automated sentiment analysis method LIWC. Emotional tone is measured by scoring negative sentiment terms (from 0 to 100) and subtracting them from a similar scoring of positive sentiment terms (from 0 to 100), resulting in a single scale from −100 (maximally negative) to 100 (maximally positive), with 0 indicating the neutral point. Error bars indicate 95% confidence intervals. Results are robust to alternative methods for measuring the emotional tone of subjects' self-reported task experience.

4 Discussion

Individuals vary substantially in how they categorize novel and controversial content, as indicated both by experimental psychological research[3–6] and by qualitative and quantitative analyses of content moderators in social media organizations[8–14]. Yet, across a variety of contexts and topics, independent groups of social media users demonstrate remarkable consistency in how they categorize controversial content[19–26], raising the question of how such consistency can arise at the collective level amid stark individual differences in classification. Building on recent work in collective intelligence[7, 33, 37–40, 64], we show that when people engage in peer-coordinated content moderation within structurally symmetrical social networks, their decisions for how to categorize and remove controversial content can become synchronized across independent communities. Importantly, our findings are a direct replication of prior models and experiments demonstrating "scale-induced convergence" in the categorization of arbitrary stimuli[7]. This work shows that the dynamics of scale enables replicable trajectories of classification to emerge despite substantial individual variation in perception; specifically, these studies show how scale can induce synchronization in classification even when no particular classification is preferred by the majority of the population. This shows how scale plays a driving role in inducing synchronization in classification, distinct from majority reinforcement.

A core strength of our study is its use of a randomized controlled design at the network-level, which enables precise identification of the effects of scale on convergence dynamics in classification. However, this design requires compromising elements of realism that mark notable limitations of our study and key areas for future research. First, our study only examines a single network topology in large teams, namely a fully-connected (or homogeneously mixing) population. This topology is the ideal starting place for

identifying the effects of scale on classification synchrony, since fully connected networks are the simplest and most minimal topology that can be held constant across scales and across populations[7], whereas more complex network features (for example, modularity or centralization) can lead to many different kinds of topologies that vary in degree along these dimensions and which are impacted by the size of the network, introducing challenges of estimation and statistical power when isolating the effects of topology. Nevertheless, an important topic for future research is to investigate how different topological structures affect coordination dynamics in the classification of controversial content. For example, consistent with prior work on social learning in forecasting tasks, more centralized networks[39] (with some people having more connections and therefore more influence than others) may enable the spread of centralized individuals' idiosyncratic interpretations, leading large independent and centralized groups to vary in their classification systems. Future research on the effects of topological structures will play a pivotal role in identifying which structures induce synchronization and which induce diversification, both of which may play valuable strategic roles in the development of classification systems depending on the context.

Another limitation of our study is its focus on only a single incentive structure. We assume the simplest and most canonical incentive structure that most directly matches classic game-theoretic models of coordination, and which is widely shown to be effective at inducing coordination in online social network experiments[65, 66]. In prior work, we show that the incentive to coordinate alone is not sufficient to induce synchronization across social groups. In Guilbeault et al.'s study[7], they found that only large, fully-connected populations develop convergent categories for novel stimuli, whereas smaller fully-connected teams generate highly variable and path-dependent categories, even though all participants in all teams are equally incentivized to coordinate and to avoid punishment from miscoordination. This demonstrates that scale and not incentive structure is the driver of structural synchronization in this setting. That says, the incentives motivating people to form classifications in the social world can vary considerably, such that an important topic for future research is to investigate how different incentive structures mediate synchronization dynamics in classification. For instance, recent work shows that a small fraction of committed defectors who refuse to coordinate (i.e., who possess a different incentive structure) can mobilize to induce tipping points that trigger the whole network to shift its choice of social convention into a classic coordination dynamic. This may be particularly relevant when considering partisan tensions and competitive dynamics in online classification processes in the context of fake news and controversial content.

In summary, our findings indicate that structural symmetry in communication dynamics can give rise to regularities in category systems at the collective level, despite considerable variation in the classification judgments among individuals. These findings inform previous studies of content moderation by showing how structured communication networks may help to ameliorate the emotional harm often experienced by workers in the field of content moderation[8, 9, 15, 61, 67]. Communication in structured social networks significantly improves workers' emotional experience and reduces psychological stress, offering an interesting direction for future research, which may explore whether networked approaches to content moderation can help to lower attrition rates among workers. This suggests that team-based content moderation practices may be preferable to current approaches involving individuals classifying content in isolation[8, 9, 68]. We anticipate structural synchronization may be further useful for an array of fields that rely on human coders to perform distributed classification tasks, from detecting various forms of misinformation[19, 20, 38] to exploring scientific data via "citizen science" (e.g., Galaxy Zoo)[59, 69], in which agreement among aggregated human judgments continues to be an essential method for the effective classification and evaluation of novel content. This network-based approach is poised to integrate into a hybrid human and AI crowdsourcing, especially given the central role that RLHF[70] (reinforcement learning from human feedback) plays in current generative AI architectures[68]. Our work suggests that the quality of human feedback into AI systems may be enriched if it is first processed through a structurally synchronized, team-based approach, which can increase the

likelihood that the human-derived classifications training AI correspond to common sense agreement and consensus. In this way, the team-based structural synchronization of classifications may enable significant improvements to the quality of classification systems not only among humans, but also among AI agents that increasingly interact with and influence human understanding throughout culture[53, 71, 72].

Acknowledgment

We gratefully acknowledge Alan Wagner for programming assistance, and also gratefully acknowledge financial support from the Content Moderation Research Award granted by Facebook. The funders played no role in the design, implementation, and write-up of this study. Finally, we are grateful to Annie Ghrist, Mariagiulia Lauro, and Paulina Paiz for research assistance.

Data Availability

All data and codes underlying this study are available as a supplement to this submission and publicly available at https://github.com/drguilbe/contentmod.

Electronic Supplementary Material

Supplementary materials including supplementary materials and methods, supplementary analyses, and supplementary references are available in the online version of this article at https://doi.org/.10.23919/JSC.2025.0024.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] P. L. Berger and T. Luckmann, *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Garden City, NY, USA: Anchor Press, 1967.

[2] R. N. Shepard and G. W. Cermak, Perceptual-cognitive explorations of a toroidal set of free-form stimuli, *Cognit. Psychol.*, vol. 4, no. 3, pp. 351–377, 1973.

[3] X. Wang and Y. Bi, Idiosyncratic tower of babel: Individual differences in word-meaning representation increase as word abstractness increases, *Psychol. Sci.*, vol. 32, no. 10, pp. 1617–1635, 2021.

[4] L. F. Barrett, *How Emotions Are Made: The Secret Life of the Brain*. London, UK: Pan Books, 2018.

[5] N. Binetti, N. Roubtsova, C. Carlisi, D. Cosker, E. Viding, and I. Mareschal, Genetic algorithms reveal profound individual differences in emotion recognition, *Proc. Natl. Acad. Sci. USA*, vol. 119, no. 45, p. e2201380119, 2022.

[6] B. Bago, M. Kovacs, J. Protzko, T. Nagy, Z. Kekecs, B. Palfi, M. Adamkovic, S. Adamus, S. Albalooshi, N. Albayrak-Aydemir, et al., Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample, *Nat. Hum. Behav.*, vol. 6, no. 6, pp. 880–895, 2022.

[7] D. Guilbeault, A. Baronchelli, and D. Centola, Experimental evidence for scale-induced category convergence across populations, *Nat. Commun.*, vol. 12, no. 1, p. 327, 2021.

[8] T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT, USA: Yale University Press, 2018.

[9] S. T. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT, USA: Yale University Press, 2019.

[10] M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, and M. S. Bernstein, The disagreement deconvolution: Bringing machine learning performance metrics in line with reality, in *Proc. 2021 CHI Conf. Human Factors in Computing Systems*, Yokohama, Japan, 2021, p. 388.

[11] B. van Aken, J. Risch, R. Krestel, and A. Löser, Challenges for toxic comment classification: An in-depth error analysis, in *Proc. 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium, 2018, pp. 33–42.

[12] N. Goyal, I. D. Kivlichan, R. Rosen, and L. Vasserman, Is your toxicity my toxicity? Exploring the impact of rater identity on toxicity annotation, *Proc. ACM Hum. Comput. Interact.*, vol. 6, no. CSCW2, p. 363, 2022.

[13] Z. Waseem, Are you a racist or am I seeing things? Annotator influence on hate speech detection on twitter, in *Proc. 1st Workshop on NLP and Computational Social Science*, Austin, TX, USA, 2016, pp. 138–142.

[14] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, Measuring the reliability of hate speech annotations: The case of the European refugee crisis, in *Proc. 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, Bochum, Germany, 2016, pp 6–9.

[15] J. A. Gallo and C. Y. Cho, *Social Media: Misinformation and Content Moderation Issues for Congress*. Washington, DC, USA: Congressional Research Service, 2021.

[16] T. Dias Oliva, Content moderation technologies: Applying human rights standards to protect freedom of expression, *Hum. Rights Law Rev.*, vol. 20, no. 4, pp. 607–640, 2020.

[17] J. Allen, C. Martel, and D. G. Rand, Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program, in *Proc. CHI Conf. Human Factors in Computing Systems*, New Orleans, LA, USA, 2022, p. 245.

[18] R. Gorwa, R. Binns, and C. Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance, *Big Data Soc.*, vol. 7, no. 1, pp. 1–15, 2020.

[19] G. Pennycook and D. G. Rand, Fighting misinformation on social media using crowdsourced judgments of news source quality, *Proc. Natl. Acad. Sci. USA*, vol. 116, no. 7, pp. 2521–2526, 2019.

[20] J. Allen, A. A. Arechar, G. Pennycook, and D. G. Rand,

Scaling up fact-checking using the wisdom of crowds, *Sci. Adv.*, vol. 7, no. 36, p. eabf4393, 2021.
[21] H. Lin, J. Lasser, S. Lewandowsky, R. Cole, A. Gully, D. G. Rand, and G. Pennycook, High level of correspondence across different news domain quality rating sets, *PNAS Nexus*, vol. 2, no. 9, p. pgad286, 2023.
[22] C. Martel, J. Allen, G. Pennycook, and D. G. Rand, Crowds can effectively identify misinformation at scale, *Perspect. Psychol. Sci.*, vol. 19, no. 2, pp. 477–488, 2024.
[23] G. Pennycook and D. G. Rand, Lack of partisan bias in the identification of fake (versus real) news, *Trends Cognit. Sci.*, vol. 25, no. 9, pp. 725–726, 2021.
[24] J. Wihbey, G. Morrow, M. Chung, and M. Peacey, *The Bipartisan Case for Labeling as A Content Moderation Method: Findings from A National Survey*. Boston, MA, USA: Ethics Institute, 2021.
[25] E. Leonardelli, S. Menini, A. P. Aprosio, M. Guerini, and S. Tonelli, Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement, in *Proc. 2021 Conf. Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, 2021, pp. 10528–10539.
[26] F. Shi, M. Teplitskiy, E. Duede, and J. A. Evans, The wisdom of polarized crowds, *Nat. Hum. Behav.*, vol. 3, no. 4, pp. 329–336, 2019.
[27] J. T. Jost, J. Glaser, A. W. Kruglanski, and F. J. Sulloway, Political conservatism as motivated social cognition, *Psychol. Bull.*, vol. 129, no. 3, pp. 339–375, 2003.
[28] P. K. Hatemi, C. Crabtree, and K. B. Smith, Ideology justifies morality: Political beliefs predict moral foundations, *Am. J. Polit. Sci.*, vol. 63, no. 4, pp. 788–806, 2019.
[29] J. M. Kivikangas, B. Fernández-Castilla, S. Järvelä, N. Ravaja, and J. E. Lönnqvist, Moral foundations and political orientation: Systematic review and meta-analysis, *Psychol. Bull.*, vol. 147, no. 1, pp. 55–94, 2021.
[30] D. Guilbeault, J. Becker, and D. Centola, Social learning and partisan bias in the interpretation of climate trends, *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 39, pp. 9714–9719, 2018.
[31] S. Garrod and G. Doherty, Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions, *Cognition*, vol. 53, no. 3, pp. 181–215, 1994.
[32] D. Guilbeault, S. Woolley, and J. Becker, Probabilistic social learning improves the public's judgments of news veracity, *PLoS One*, vol. 16, no. 3, p. e0247487, 2021.
[33] J. Becker, D. Brackbill, and D. Centola, Network dynamics of social influence in the wisdom of crowds, *Proc. Natl. Acad. Sci. USA*, vol. 114, no. 26, pp. E5070–E5076, 2017.
[34] D. Brackbill and D. Centola, Impact of network structure on collective learning: An experimental study in a data science competition, *PLoS One*, vol. 15, no. 9, p. e0237978, 2020.
[35] D. Guilbeault and D. Centola, Networked collective intelligence improves dissemination of scientific information regarding smoking risks, *PLoS One*, vol. 15, no. 2, p. e0227813, 2020.
[36] J. A. Becker, D. Guilbeault, and E. B. Smith, The crowd classification problem: Social dynamics of binary-choice accuracy, *Manage. Sci.*, vol. 68, no. 5, pp. 3949–3965, 2021.
[37] L. Hong and S. E. Page, Groups of diverse problem solvers can outperform groups of high-ability problem solvers, *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 46, pp. 16385–16389, 2004.
[38] Z. Kunda, The case for motivated reasoning, *Psychol. Bull.*, vol. 108, no. 3, pp. 480–498, 1990.
[39] E. J. Horberg, C. Oveis, D. Keltner, and A. B. Cohen, Disgust and the moralization of purity, *J. Pers. Soc. Psychol.*, vol. 97, no. 6, pp. 963–976, 2009.
[40] M. Dehghani, K. Johnson, J. Hoover, E. Sagi, J. Garten, N. J. Parmar, S. Vaisey, R. Iliev, and J. Graham, Purity homophily in social networks, *J. Exp. Psychol. Gen.*, vol. 145, no. 3, pp. 366–375, 2016.
[41] A. Boutyline and S. Vaisey, Belief network analysis: A relational approach to understanding the structure of attitudes, *Am. J. Sociol.*, vol. 122, no. 5, pp. 1371–1447, 2017.
[42] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. B. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky, Exposure to opposing views on social media can increase political polarization, *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 37, pp. 9216–9221, 2018.
[43] J. J. Van Bavel and A. Pereira, The partisan brain: An identity-based model of political belief, *Trends Cognit. Sci.*, vol. 22, no. 3, pp. 213–224, 2018.
[44] J. J. Van Bavel, S. Rathje, M. Vlasceanu, and C. Pretus, Updating the identity-based model of belief: From false belief to the spread of misinformation, *Curr. Opin. Psychol.*, vol. 56, p. 101787, 2024.
[45] S. J. Taylor, L. Muchnik, M. Kumar, and S. Aral, Identity effects in social media, *Nat. Hum. Behav.*, vol. 7, no. 1, pp. 27–37, 2023.
[46] C. R. Sunstein, The law of group polarization, *J. Polit. Philos.*, vol. 10, no. 2, pp. 175–195, 2002.
[47] C. R. Sunstein, *Going to Extremes: How Like Minds Unite and Divide*. New York, NY, USA: Oxford University Press, 2011.
[48] N. Epley and T. Gilovich, The mechanics of motivated reasoning, *J. Econ. Perspect.*, vol. 30, no. 3, pp. 133–140, 2016.
[49] H. H. Nam, J. T. Jost, and J. J. Van Bavel, "Not for all the tea in China!" political ideology and the avoidance of dissonance-arousing situations, *PLoS One*, vol. 8, no. 4, p. e59837, 2013.
[50] E. Jahani, N. Gallagher, F. Merhout, N. Cavalli, D. Guilbeault, Y. Leng, and C. A. Bail, An Online experiment during the 2020 US–Iran crisis shows that exposure to common enemies can increase political polarization, *Sci. Rep.*, vol. 12, no. 1, p. 19304, 2022.
[51] J. Becker, E. Porter, and D. Centola, The wisdom of partisan crowds, *Proc. Natl. Acad. Sci. USA*, vol. 116, no. 22, pp. 10717–10722, 2019.
[52] D. Guilbeault, S. Delecourt, T. Hull, B. S. Desikan, M. Chu, and E. Nadler, Online images amplify gender bias, *Nature*, vol. 626, no. 8001, pp. 1049–1055, 2024.
[53] D. Guilbeault, S. Delecourt, and B. S. Desikan, Age and gender distortion in online media and large language models, *Nature*, vol. 646, no. 8087, pp. 1129–1137, 2025.
[54] D. Guilbeault, E. O. Nadler, M. Chu, D. R. Lo Sardo, A. A. Kar, and B. S. Desikan, Color associations in abstract

semantic domains, *Cognition*, vol. 201, p. 104306, 2020.
[55] A. Baronchelli, T. Gong, A. Puglisi, and V. Loreto, Modeling the emergence of universality in color naming patterns, *Proc. Natl. Acad. Sci. USA*, vol. 107, no. 6, pp. 2403–2407, 2010.
[56] Y. R. Tausczik and J. W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, *J. Lang. Soc. Psychol.*, vol. 29, no. 1, pp. 24–54, 2010.
[57] R. Hastie and T. Kameda, The robust beauty of majority rules in group decisions, *Psychol. Rev.*, vol. 112, no. 2, pp. 494–508, 2005.
[58] M. Amraoui, T. B. Stambouli, and B. Alshaqaqi, On using the wisdom of the crowd principles in classification, application on breast cancer diagnosis and prognosis, *Int. J. Bioinf. Res. Appl.*, vol. 15, no. 4, pp. 324–346, 2019.
[59] S. Coughlin, S. Bahaadini, N. Rohani, M. Zevin, O. Patane, M. Harandi, C. Jackson, V. Noroozi, S. Allen, J. Areeda, et al., Classifying the unknown: Discovering novel gravitational-wave detector glitches using similarity learning, *Phys. Rev. D*, vol. 99, no. 8, p. 082002, 2019.
[60] K. Crowston, Gravity spy: Humans, machines and the future of citizen science, in *Proc. 2017 ACM Conf. Computer Supported Cooperative Work and Social Computing*, Portland, OR, USA, 2017, pp. 163–166.
[61] E. Dwoskin, *Facebook Content Moderator Details Trauma That Prompted Fight for $52 Million PTSD Settlement*, http://www.shturl.cc/53edc56bc591a8b65376b9cd7c4cc121, 2018.
[62] A. Hern, *Revealed: Catastrophic Effects of Working as A Facebook Moderator*, http://www.shturl.cc/d969c56a3aa3cc9807fa185d9bb3b659, 2019.
[63] S. B. Srivastava, A. Goldberg, V. G. Manian, and C. Potts, Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations, *Manage. Sci.*, vol. 64, no. 3, pp. 1348–1364, 2017.
[64] D. Centola, D. Guilbeault, U. Sarkar, E. Khoong, and J. Zhang, The reduction of race and gender bias in clinical treatment recommendations using clinician peer networks in an experimental setting, *Nat. Commun.*, vol. 12, no. 1, p. 6585, 2021.
[65] D. Centola and A. Baronchelli, The spontaneous emergence of conventions: An experimental study of cultural evolution, *Proc. Natl. Acad. Sci. USA*, vol. 112, no. 7, pp. 1989–1994, 2015.
[66] D. Centola, J. Becker, D. Brackbill, and A. Baronchelli, Experimental evidence for tipping points in social convention, *Science*, vol. 360, no. 6393, pp. 1116–1119, 2018.
[67] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, Deeper attention to abusive user content moderation, in *Proc. 2017 Conf. Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017, pp. 1125–1135.
[68] K. Hao, *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*. New York, NY, USA: Penguin Press, 2025.
[69] D. Watson and L. Floridi, Crowdsourced science: Sociotechnical epistemology in the e-research paradigm, *Synthese*, vol. 195, no. 2, pp. 741–764, 2018.
[70] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, Fine-tuning language models from human preferences, arXiv preprint arXiv: 1909.08593, 2019.
[71] L. Brinkmann, F. Baumann, J. F. Bonnefon, M. Derex, T. F. Müller, A. M. Nussberger, A. Czaplicka, A. Acerbi, T. L. Griffiths, J. Henrich, et al., Machine culture, *Nat. Hum. Behav.*, vol. 7, no. 11, pp. 1855–1868, 2023.
[72] P. E. Smaldino, A. Russell, M. R. Zefferman, J. Donath, J. G. Foster, D. Guilbeault, M. Hilbert, E. A. Hobson, K. Lerman, H. Miton, et al., Information architectures: A framework for understanding socio-technical systems, *npj Complex.*, vol. 2, no. 1, p. 13, 2025.

Damon Centola received the PhD degree in sociology from Cornell University, NY, USA. Today, he is the Elihu Katz Professor of communication, sociology, and engineering at University of Pennsylvania, PA, USA, where he directs the Network Dynamics Group. He received the American Sociological Association's Award for Outstanding Research in Mathematical Sociology in 2006, 2009, and 2011; the Goodman Prize for Outstanding Contribution to Sociological Methodology in 2011; the James Coleman Award for Outstanding Research in Rationality and Society in 2017; and the Harrison White Award for Outstanding Scholarly Book in 2019. His research is funded by the National Science Foundation, the Robert Wood Johnson Foundation, the National Institutes of Health, the James S. McDonnell Foundation, Facebook, and the Hewlett Foundation. He is a series editor for Princeton University Press, and the author of *How Behavior Spreads: The Science of Complex Contagions* (2018) and *Change: How to Make Big Things Happen* (2021).

Douglas Guilbeault received the PhD degree from University of Pennsylvania, PA, USA. After his PhD, he was an assistant professor in the management of organizations at Berkeley's Haas School of Business, University of California Berkeley, USA, and today he is an assistant professor of organizational behavior at Graduate School of Business, Stanford University, USA. He co-directs the Computational Culture Lab, which harnesses and builds computationally intensive network- and language-based methods to study how organizational cultures emerge and evolve. His work has appeared in a number of top journals, including *Nature* and *Management Science*, as well as in popular news outlets, such as *The Atlantic* and *The Harvard Business Review*. He has received top research awards from International Conference on Computational Social Science, Cognitive Science Society, and International Communication Association.