Neighborhood topology shapes narrative interaction dynamics in networked groups

J. Hunter Priniski[1,*], Bryce Linford[1], Anna Hirschmann[2], Sai Krishna[3], Fred Morstatter[3], Nancy Rodriguez[2], Jeff Brantingham[4], and Hongjing Lu[1,5]

[1]Department of Psychology, University of California, Los Angeles
[2]Department of Applied Mathematics, University of Colorado
[3]Information Sciences Institute, University of Southern California
[4]Department of Anthropology, University of California, Los Angeles
[5]Department of Statistics, University of California, Los Angeles
[*]priniski@ucla.edu

ABSTRACT

Introduction

Since the dawn of human culture, people have engaged in discussions of real or imagined events, collectively interpreting them in terms of narratives. Today, digital media and online communication networks have dramatically impacted the way narratives are formed, taking shape through formats such as hashtags, tweets, and other digital expressions. Through narrative communications, online networks influence individual beliefs[1,2] and group consensus-making[3–5], with real-world impact on political polarization[6] and collective organizing in both online and offline settings[2,7–9]. Empirically investigating networked behavior is a challenging task because the narratives that arise in online contexts tend to be sprawling and unwieldy[10], and modern social media environments are not amenable to experimental control.

To better understand how online communication impacts narrative adoption and belief change, we report an experiment on networked groups of individuals incentivized to align narrative-based interactions with network neighbors when generating hashtags to characterize a disaster event. Hashtags are a distinctive marker of narrative interaction on social media. They function as concise representations of complex narratives[11–13], and connect spatially disorganized groups according to the content of their shared narratives and goals, thus constituting a potent force for online activism[9,14,15]. Across an online network, hashtags serve as topic labels for generated content, which assist online platforms with the algorithmic categorization, curation, and dissemination of dynamic social media discourse[16] (e.g., by mapping discrete units of online content shared over time to a single event or discussion)[9,14]. At an individual level, hashtags allow users to signal personal contributions to broader narratives emerging from interactions in an online community. Previous research on hashtags has primarily focused on understanding their linguistic and semantic content[17] and modeling the dynamics of their adoption and online spread in real-world (i.e., scale free) social media networks. For example, mathematical models suggest that dominant (i.e., widely shared) hashtags emerge as a result of a preferential attachment mechanism that increases the popularity of early popular hashtags over time. While many hashtags initially compete for popularity, only a small set of hashtags persist to allow for broader narrative collaboration across the network[18,19]. It remains unclear how an individual's prior knowledge interacts with the structure of their online network (e.g., network diameter or neighborhood size) to influence hashtag generation, and how exposure to hashtags generated by others within their networked group shapes beliefs and the subsequent generation of related narratives.

We approach the study of narrative interaction by examining language generation and decision-making in online network experiments. In the experiments, groups of participants are placed in an online social network and interact with each other with the goal of receiving rewards for producing behaviors that align with network neighbors. This paradigm allows us to manipulate the network structure and complexity of content in online communication, potentially providing insights into the dynamics of both individual- and group-level phenomena. Researchers have previously used relatively simple materials to investigate how varying social network structure (e.g., node connectivity) influences the adoption of coordinated behaviors. For example, Centola and Baronchelli (2015) introduced the *Name Game* asking participants to coordinate with network neighbors on a face-naming task[4]. If network neighbors provided the same name as a response for a presented image of a face, they were awarded 50 cents; otherwise they lost 25 cents. The researchers found that group consensus can result from purely local

incentives to interacting network neighbors. Furthermore, network structure impacts the dynamics of consensus formation. Specifically, interactions within *homogeneously-mixed* or fully-connected networks, where each participant can potentially interact with any other member of the network, exhibited a rapid convergence to consensus (i.e., the full network aligning on a single name). In contrast, interactions within *spatially-embedded networks*, where each participant is linked to only a handful of close-by neighbors, did not show strong convergence (see Figure 7).

In the present study, we extend this experimental paradigm by using naturalistic narrative materials (a description of a real-world disaster with inherent causal structure), coupled with a type of networked interaction behavior (generating hashtags) that occurs in real-world online settings. Rather than simply documenting the impact of varying network structure on the dynamics of consensus formation, we examine how the network communications impacts decision strategies adopted by individuals over time, and in turn individual decisions shape the collective behavior of the group through online communications. We hypothesize that network structures shape the information that individuals gather from their social context. Individuals, in turn, make rational decisions by integrating this socially gathered information through online communications with their own prior knowledge to generate personalized narratives. Group-level consensus emerges from these individual decisions, even as individuals are just rewarded based on successful when local coordination with network neighbors. To test this hypothesis, we collected a rich set of empirical data through an online experimental platform, spanning multiple network sizes, two distinct network structures, and two types of interaction tasks. We compared human performance with an agent-based model—Context Aware Agents (CAA), which integrates prior knowledge with social context to facilitate effective local coordination. Our goal was to assess whether the model accurately captures the dynamics of human consensus formation across all experimental conditions. In addition to using hashtags as a narrative format, we asked participants to generate other forms of narratives, such as tweets, describing the same event before and after networked interactions. We then examined whether different network structures influenced shifts in personalized narratives, particularly in terms of increased causal content, when reflecting on a complex real-world disaster.

Results

Figure 1 provides an overview of the experiment. Participants read a narrative-based passage describing the Fukushima nuclear disaster and its effects on local communities and the environment, and generated a personal narrative and hashtags describing the event. Participants were assigned to either the experimental narrative interaction group, which generated hashtags during networked interactions involving narratives, or the control group, which performed a face-naming task in the Name Game from Centola and Baronchelli (2015) as their networked interaction activity. We also manipulated network structures, including homogeneously-mixed or spatially-embedded networks. Participants were financially incentivized to coordinate responses with network neighbors. Following online communication, participants wrote another personal narrative and hashtags about the Fukushima disaster.

Coherence dynamics during networked narrative interaction

We analyzed 41,600 interactions collected from 1040 participants across 26 experimental runs. Here, we focus on how the form of networked interaction (narrative interaction via hashtag-matching versus playing the Name Game as a control reference) mediates the impact of neighborhood structure (homogeneous versus spatial) on the emergence of group-level behavioral coherence. Because matching hashtags requires coordinating on specific causal and/or semantic content in the narrative, whereas naming an image of a face does not, hashtag coordination can be considered as a more computationally complex interaction task. By comparing hashtag-matching dynamics against face-naming as a control, we can assess how the increased computational complexity of aligning behaviors based on an array of causal and semantic information (i.e., narrative information) impacts group-level outcomes. In addition, we examined two types of network coherence: *group-level coherence*, the similarity of responses across all participants in a group, and *local coherence*, the similarity of responses between interacting nodes (i.e., network neighbors along network edges). We present results from both coherence measures.

Interaction task shapes onset of behavioral coherence within networked groups

We examined two measures of group-level coherence: (1) the proportion of a group reporting a shared, or *normative* response, and (2) the entropy of the full response distribution of a group[20]. When assessing how network structures impact behavioral dynamics, researchers have generally predicted the proportion of a network producing a dominantly shared behavior at a given time[3–5]. A shortcoming of this winner-take-all measure is that it does not capture the variability of other alternative non-dominant responses. The full set of responses from a group compose a distribution of behaviors that can be heavy-tailed, multi-modal, or highly skewed — key distributional features not captured by simply predicting the modal response. Hence, we also analyzed the results based on the entropy of the response distribution of a group. The entropy metric provides a concise measure of response variability across the entire group: lower entropy indicates more similar or coherent responses in the group, while higher entropy suggests greater diversity or variation in responses[20,21].

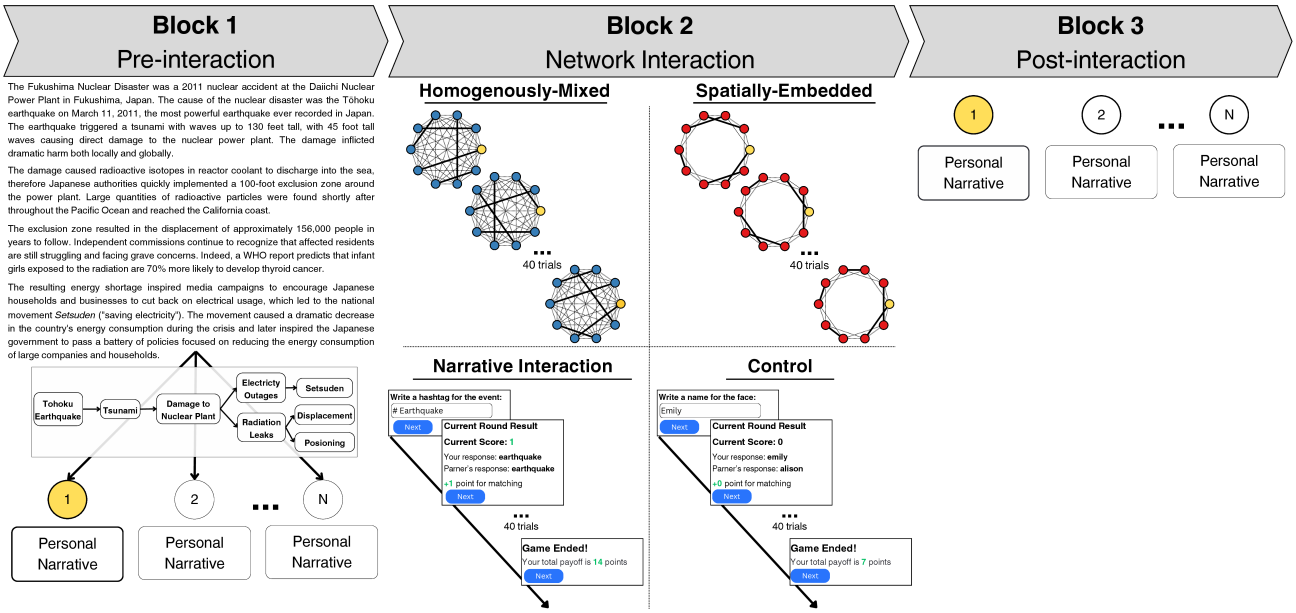

Figure 1. Experiment procedure and networked interaction tasks. The experimental design follows three blocks. We highlight a single node (in yellow) to illustrate a single participant's tasks through the procedure. In the pre-interaction block, all participants read the Fukushima nuclear disaster narrative that encodes the graphical causal model illustrated. The causal model was not presented to participants. Participants then wrote a tweet-like personal narrative about the disaster and generated ten hashtags describing the event. Participants next entered a network interaction block where they communicated with network neighbors. In the network interaction block, group communication varied as a function of network structure (homogeneously-mixed vs spatially-embedded) and content of interaction (narrative interaction based on hashtag matching vs control based on the Name Game from Centola and Baronchelli (2015)). Participants interacted with neighbors, randomly chosen based on network structure, for 40 trials and received one point for each trial in which their response matched their neighbor's (the participant with the most points at the end of network interaction received a financial reward). In the post-interaction block, participants wrote a personal narrative about the Fukushima nuclear disaster and ten more hashtags describing the event.

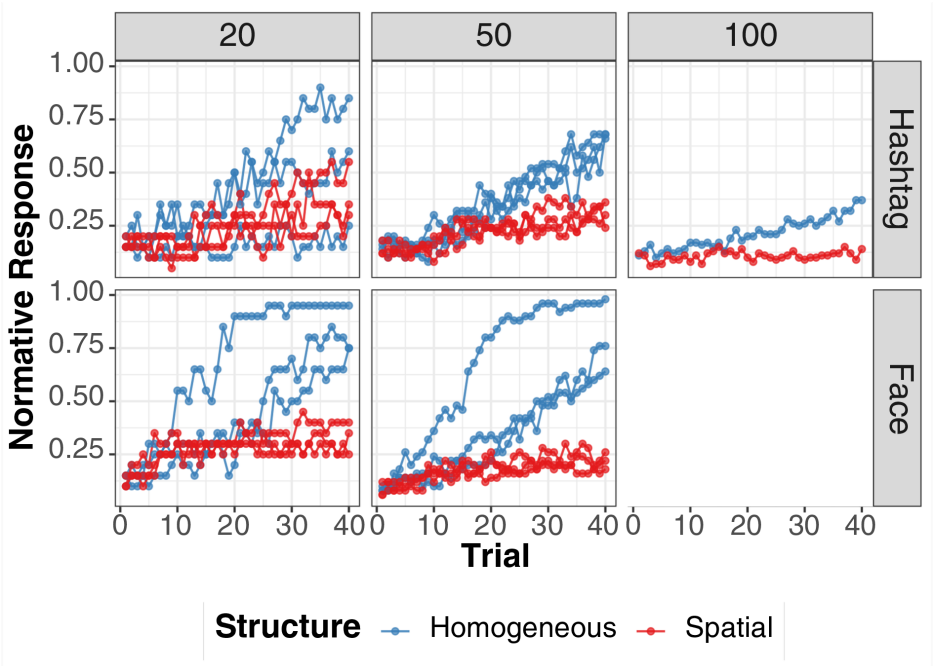

Figure 2. Onset of behavioral coherence during networked interaction. Panels display the proportion of each group adopting a dominant response over course of interaction by group size (columns) and interaction media content (rows). Each line represents a single experimental run from a group of participants, and each point represents the proportion of the group reporting the most common response (which can change trial to trial for a given experimental run). The entropy results from the experimental runs are in the supplemental materials SM 2.

.

We examine response convergence in networked interactions using two metrics: proportion of a group providing the dominant,'normative' response and the entropy of each group's response distribution. We fit a Beta-distributed GLM to predict the proportion of a group producing a dominant response as a function of *trial number*, neighborhood *structure*, *content* of network interaction, and their interactions, while controlling for network size. As shown in Figure 2, shared behaviors emerged reliably over time in homogeneously-mixed face-naming networks ($\beta_{Trial} = 0.09$, 95% CI $[0.09, 0.10]$), doing so more slowly in homogeneously-mixed hashtag-matching networks ($\beta_{Trial:Hashtag} = -0.04$, 95% CI $[-0.05, -0.03]$), and substantially less so in spatially-embedded face-naming networks ($\beta_{Trial:Spatial} = -0.07$, 95% CI $[-0.08, -0.07]$). These findings are consistent with previous research analyzing group communication over face naming[4]. In addition, we found a significant three-way interaction effect between the network structure, the content of the interaction, and the number of trials, such that the rate at which shared behaviors emerge across network structures is shaped by the complexity of interaction($\beta_{Trial:Spatial:Hashtag} = 0.05$, 95% CI $[0.04, 0.06]$). These findings are qualitatively consistent with simulation results that suggest that the cognitive complexity (e.g., increased partner response uncertainty) of the interaction task can impact the rate of consensus[22,23].

Next, we fit a Gaussian-distributed GLM to predict the change in the entropy of the full response distribution across experimental runs. As shown in Supplemental Figure SM 2, a group's response entropy steadily decreased as a function of subsequent interactions in homogeneously-connected face-naming networks ($\beta_{Trial} = -0.06$, 95% CI $[-0.07, -0.06]$), doing so more slowly in spatially-embedded face-naming networks ($\beta_{Trial:Spatial} = 0.04$, 95% CI $[0.03, 0.04]$). Entropy also decreased more slowly in groups matching hashtags ($\beta_{Trial:Hashtag} = 0.02$, 95% CI $[0.03, 0.02]$), suggesting that coherence emerges more slowly in situations with increased cognitive complexity. Furthermore, the analysis revealed a significant three-way interaction between network structure, interaction content, and trial number, such that the rate of entropy change across network structures is larger in face-naming networks than hashtag-matching networks ($\beta_{Trial:Spatial:Hashtag} = -0.02$, 95% CI $[-0.03, -0.01]$). This finding suggests that increasing the cognitive complexity of interaction tasks can reduce the effect of a network's structure on group-level coherence dynamics, such as homogeneously-mixed networks reaching consensus.

The role of individual decision strategies in the formation of group consensus

We analyzed individual-level decision strategies in networked interactions as a function of network structure and interaction content. We found that human participants explore longer by sampling new responses when performing high complexity interaction task than low complexity interactions. For tasks with low complexity (e.g., face naming), individuals are more likely

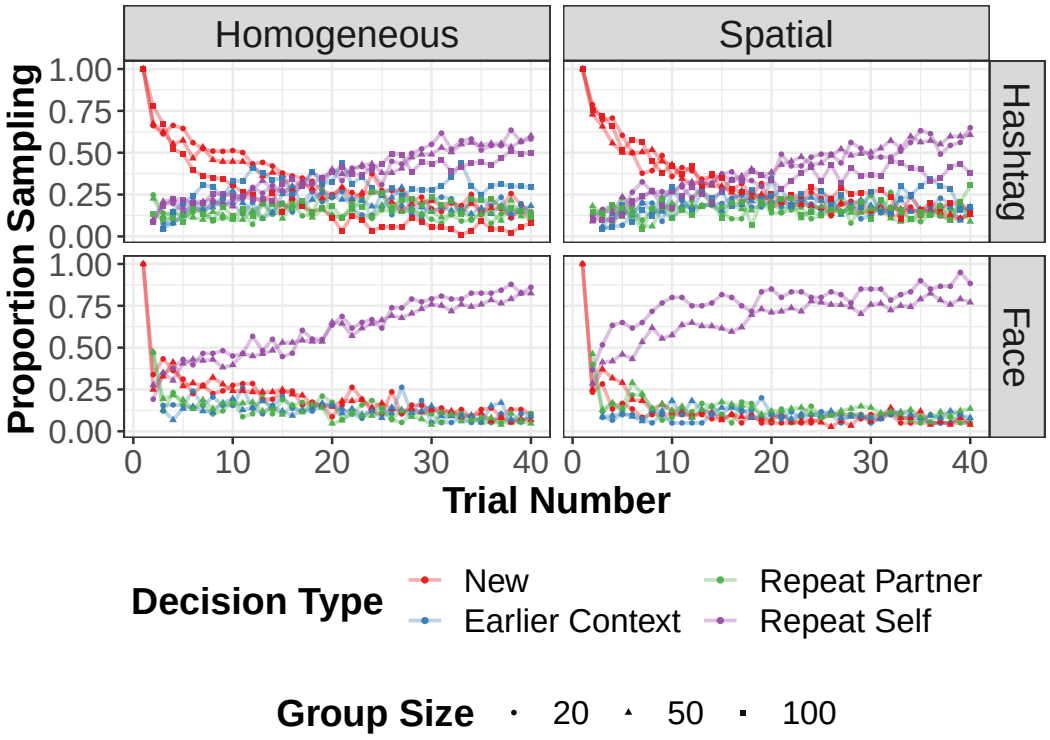

Figure 3. Dynamics of individual decision strategy during networked interaction. Each panel illustrates the temporal dynamics of the proportion of each group adopting one of the four decision strategies (sampling new responses, repeating a partner's last response, repeating one's own previous response, and resampling from earlier context) across 40 trials in different network structures (Spatially-embedded, Homogeneously-mixed) and interaction contents (Hashtag-matching, Face-naming). We code cases when both a partner and oneself produces the same response (i.e., a rewarded response) as repeat self. Each panel displays the proportion of participants choosing a particular strategy as a function of trial number, differentiated by group size (20, 50, 100). Displayed proportions are averaged across runs for a given size, structure, and content. As shown in the top panels, hashtag-matching networks sample new responses (red lines) for much longer than face-naming groups (bottom panels). In contrast, participants in face-naming group are more quick to adopt the self-consistent strategy (repeating oneself, depicted by the purple lines).

to generate self-consistent responses that mirror social context. This divergence in individual-level strategies helps explain why interaction content mediates the effect of network structure on group-level consensus.

On a given trial, participants weighed between several decision strategies to make a response by exploring new responses from their prior knowledge and exploiting responses acquired from social context. We consider four types of decision strategies that a participant could have adopted while interacting with network neighbors. They could explore the response space by sampling a "new" response that neither a partner nor they themselves had produced on a previous trial. They could copy their partner's response from the last trial, or repeat their previous response ("repeat partner" and "repeat self," respectively). In cases where there is agreement on a previous trial (i.e., repeat partner and repeat self are the same response), we code this as repeat self, as self-consistency is rewarded in these trials. They could also resample a response they remember from earlier interactions ("earlier context"), which could have been generated or received in a previous networked interaction trial.

As shown in Figure 3, we observed that increasing the cognitive complexity of interaction (i.e., hashtag matching versus face naming) encourages groups to explore new responses for longer. Regardless of network structure and size, the number of participants who sampled new responses gradually decreased over time in hashtag-matching networks (top panels), The number who sampled new responses decreased quickly in face-naming networks (bottom panels). In hashtag-matching networks, the preference for sampling new responses from background knowledge for longer makes hashtag interaction a more cognitively complex task because it is more difficult to predict which responses are most likely to produce a coordination reward[22,24]. It also slows the network's ability to reach consensus. As a corollary, groups exploring new responses for longer reduces the influence of network structure on group's potential to adopt shared behaviors. Indeed, the effect of network structure on group-level outcomes is more pronounced in low-complexity interactions of the face naming task, where mirroring social

context more reliably produces matching behaviors with network neighbors.

But why do hashtag-matching networks explore new responses longer than face-naming networks? We hypothesize that the cognitive complexity of the interaction task plays an important role here as well. First, as shown in Figure SM 4, hashtags generated during network communication often align with one of the discrete events described by the narrative (e.g., #Earthquake, #Tsunami, and #Setsuden, each resulting in coordination in about $1/3$ of the trials they're generated). Second, hashtags that align with the same underlying narrative entity (e.g., a specific event or topic) can have different string representations, a phenomenon Booten (2016) describes as individualistic hashtags in scale-free, real-world online networks[17]. To test this hypothesis, for each response we measured the probability that a participant coordinated with their partner given they generate that response across all experimental runs; we refer to this value as the *coordination rate* of the response. #Nuclear, #RememberFukushima, and #NuclearDisaster are the three hashtags with the highest coordination rates across all experimental runs, with probabilities that a participant coordinates when generating those hashtags ranging between between 33% and 40%. Each of these hashtags encodes broad topic-level information about the disaster. However, doing so with different strings makes coordination more difficult (i.e., a factor that increases the cognitive complexity of the hashtag-based interaction task) . When a participant considers which hashtag to generate on a given trial, they not only need to first consider which causal/semantic entity their partner may focus on but also predict which string representation they will choose for encoding that discrete entity. This variation in content amplifies the cognitive demands of the interaction task by expanding the potential coordination space beyond exact narrative entity alignment. Because the response a neighbor may generate on a given trial is less certain, mirroring one's social context (repeating their own or a previous partner's response) becomes a lower utility strategy.

We also examined the coordination rates of the face responses. As shown in Figure SM 4, the coordination rates for responses in the face-naming groups, with the most successful names—Emily (coordination .65), Maddie (.6), and Taylor (.6)—resulting in coordination on over half the trials in which they are generated. This is much higher than the top hashtag coordination rates, which top out at around .4. Indeed, ten names have higher average coordination rates than the top hashtag. Participants are more willing to adopt names received in social context than hashtags, as there is less background knowledge shaping which responses are viable. This is observed in Figure SM 1 by comparing the initial distributions for names and hashtags in the first round of interactions. Hashtags have a more skewed, less uniform distribution than face names. Due to interaction dynamics, face names become less uniform than hashtags as shared responses are reached in the face naming condition. The lack of causal and semantic complexity in the face naming task, as compared to generating hashtags for the disaster narrative, allows social context to play a more important role in sampling face responses on a given trial, letting social context seep in to guide predicted coordination utilities. Other participants' responses encountered in social context are more readily incorporated into participant's sampling strategies. This is in contrast to the hashtag interaction conditions where responses can also align on an array of causal/semantic relations, making different responses less interchangeable. For example, Mary and Emily are effectively neutral and interchangeable responses given the lack of causal content communicated by the face stimuli, whereas #NaturalDisaster and #Tsunami are less interchangeable as they isolate different causal relations in the narrative materials. Therefore, as observed by the higher coordination rates among face names than hashtag responses, face naming groups quickly reach consensus because responses encountered in social context can be more easily adopted and result in coordination rewards than hashtag responses.

Shifts in narrative causal framing following networked interactions

Our current findings show that rewards for coordinating hashtags about narrative content in networked interactions can facilitate shared behaviors (e.g., dominant hashtag responses), but do networked interactions impact the narratives people generate as well? Before and after network interaction, participants wrote tweet-like personal narratives about the Fukushima nuclear disaster. We used natural language processing (NLP) methods to analyze the narratives generated by participants. Numerous studies have shown that causal relations are central to narrative representation[25–28]. Hence, we focus on analyzing the causal claims that participants made in their written narratives.

Priniski et al. (2023) developed a causal language identification model that identifies and extracts causal claims expressed in text documents[29]. The model identifies spans of words that serve as input to explicitly stated causal relations. Both the cause and effect events, and the underlying causal relation (i.e., a causal trigger) is explicitly stated for the algorithm as prior knowledge to identify the causal claim. The extracted claims are then co-referenced based on embeddings of the identified entities as computed by a fine-tuned RoBERTa-XL transformer model[30,31], to produce clusters of semantically similar topics, termed "causal topics". The model additionally encodes the direction of the stated causal relationships linking any two topics. To extract causal relationships expressed in the personal narratives, we used the causal language model to analyze all personal narratives generated by participants before and after networked interaction. The model identified documents expressing explicit causal claims (i.e., a cause-effect relationship), and clustered the claims based on their semantic content. As shown in Supplemental Table SM 1 and Figure SM 5, the model identified 20 distinct causal topics, with topics relating to the events described in the narrative (e.g., Earthquake, Tsunami, Nuclear Disaster), in addition to broad semantic-level topics not explicitly expressed in

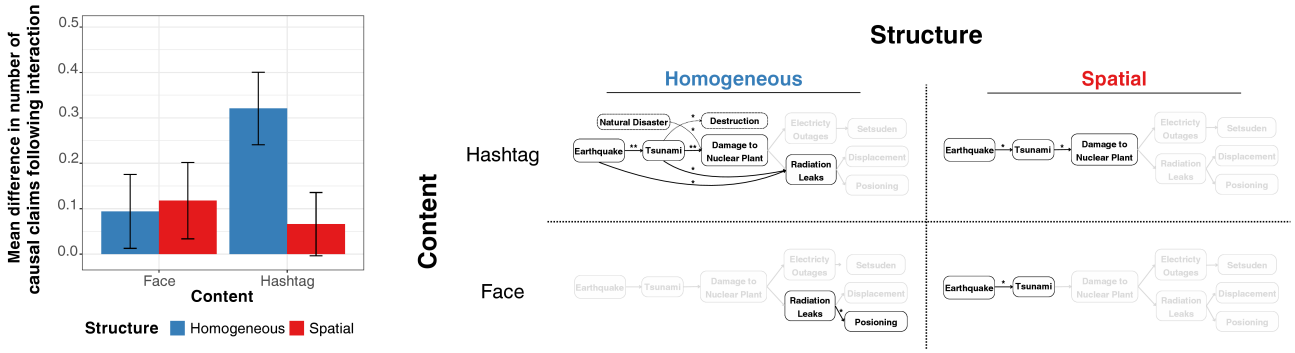

Figure 4. Shifts of causal claims in personal narratives following networked interaction. Left: the mean difference scores in the number of causal claims per participant under each of the interaction conditions. Both content conditions engaged in the same pre- and post-interaction phases (see Figure 1). Participants in homogeneously-mixed hashtag networks exhibit the largest increase in the amount of causal language generated following networked interaction. **Right:** Of those who demonstrated a change in the number of causal claims after networked interaction, we computed a one-sample t-test on the difference scores (the number of causal claims following networked interaction − number of causal claims before interaction for each participant) to measure significant shifts in expressed causal relations relations. The causal language model identified twenty distinct topics, which mapped onto each of the discrete narrative entities shown in solid boxes in the causal diagrams, in addition to broad semantic topics shown with dashed lines. Each drawn line shows a causal relationship expressed significantly more following networked interaction. Participants in the both hashtag interaction condition shifted around the complete generative causal chain in the narrative, while those in the face-naming condition did not. Furthermore, homogeneously-mixed hashtag-matching networks showed significant shifts for more causal relations than the other groups, suggesting that participants in these networks are exposed to a variety of causally-related hashtags that dramatically shift their personal narratives. More details about the causal relations with significant shifts can be found in supplemental Table SM 2.

the narrative (i.e., a topic for Natural Disaster which refers to the earthquake/tsunami in the narrative). Each document with an identified causal relation received a cause cluster label and an effect cluster label, with some clusters relating more to causes and some more to effects. As shown in Supplemental Table SM 1, to reduce noise in the statistical analysis of causal language shift, we remapped the unsupervised topic labels onto the Fukushima narrative's casual model (see Figure 6), collapsing redundant topics as necessary. Specifically, the causal language model identified twenty distinct topics, which mapped onto each of the eight discrete narrative entities shown in the causal diagrams, in addition to broad semantic topics shown with dashed lines.

As shown in the left panel of Figure 4, we conducted an independent sample t-test on the difference scores (i.e., number of causal claims generated after interaction − number of causal claims generated before interaction for each participant) across both levels of structure and content. The homogeneously-connected structure with hashtag-matching was the only group that indicated an average difference score statistically significant different from zero ($t(261) = 4.01$, $p < .001$). Neither hashtag spatial ($t(257) = 0.95$, $p = .345$), nor the face-naming networks had a significant shift (homogeneous difference values, $t(201) = 1.16$, $p = .249$); spatial: $t(203) = 1.40$, $p = .164$). Supplemental Figure SM 6 shows the distribution of difference scores in each of the network structure and interaction content conditions. To ensure that this effect is robust we additionally fit a Gaussian hurdle model to the distribution of difference scores as a function of network interaction as linear predictors (structure and interaction content) (see Supplemental Information). We found converging evidence that the hashtag interaction yield a significant effect on change in causal language in personal narratives, with the largest shift among homogeneously-mixed networks.

We now narrow our analysis of causal language shifts based on the subset of participants for who we observed a change in causal claims after networked interaction (i.e., among the approximately 50% of participants who had a shift in their personal narratives). We analyzed which specific causal relationships increased (i.e., more participants mentioned that causal relation after interaction) and decreased after networked interaction, to assess how interactions shifted how participants wrote about the nuclear disaster narrative's contents. We performed a one-sample t-test to assess which causal relationships expressed a positive or negative shift for each of the network conditions by computing the subject-level difference scores (post-interaction count − pre-interaction count) for each causal relationship and comparing the mean of that distribution to 0.0 (null hypothesis, no shift). The differences that are significantly different from 0.0 resulted in a list of shifted causal relations in post-interaction personal narratives, highlighting which causal relationships were more pronounced in participants' written documents following network

interaction. The right component of Figure 4 shows which causal relations had post-interaction shifts significantly greater than zero. A complete list of significant shifts is provided in supplemental Table SM 2, and shift statistics for each of the 20x20 causal relations identified by the pipeline in this project's repository on the Open Science Framework. As shown in the top panels on the right part of Figure 4, for both of the hashtag-matching groups, causal language change centered around the three causal events (Earthquake, Tsunami, and Nuclear Disaster) describing the generative causal chain in the narrative. Furthermore, homogeneously-mixed hashtag-matching interactions resulted in more participants in the group eliciting the full causal chain in the narrative ($p < .001$) in addition to an array of additional causal relations. The spatially-embedded hashtag-matching networks only showed a shift towards the initial generative causal chain ($p < .05$). The results are different for the face-naming groups, where the causal chain was not significant in either network structure condition. These results suggest that wider exposure to the causal and semantic content of a full group of participants' hashtags (i.e., as is the case for individuals in homogeneously-mixed hashtag-matching groups) has a substantive impact on the causal content that they reference in their written documents about the event. With this network interaction effect on causal language being less pronounced in more insular, spatially-embedded hashtag-matching neighborhoods, and approximately no impact when groups coordinate about causally-irrelevant materials to the narrative (i.e., when coordinating on naming a face rather than writing hashtags for the narrative).

Simulating behavioral dynamics with networks of *Context Aware Agents (CAA)*

Game theory offers a natural framework for modeling decision-making in scenarios where the actions of others influence the outcome for each participant[32,33]. A well-known example is the replicator equation, frequently applied in evolutionary game theory to model how the proportion of different strategies in a population shifts over time, driven by their comparative success[34]. Strategies yielding higher-than-average payoffs tend to increase in a population, while less successful ones diminish, fostering the rise of dominant strategies. However, these models assume predefined outcomes solely based on each player's strategies, which does not align with the dynamics of our experiment or real-world online networks. In this experiment, the choice to adopt one strategy over another may or may not be a good choice depending on if responses happened to align with others in the network. For this reason, it is difficult to determine a payoff matrix for the strategies of this experiment, requiring different modeling frameworks (i.e., agent-based modeling). We should note that if we make the "name" or "hashtag" the strategies then the payoff matrix would simply be the identity matrix, but we do not gain much understanding from such model.

Behavioral experiments have shown that decision-making heuristics are influenced by a combination of contextual factors and prior knowledge[35–37]. Here, we propose and test an agent-based model, where individuals update their strategies by integrating prior knowledge with new contextual insights during network interactions over the course of the experiment, representations we term as 'social context'. In our experiments, decision-making strategies consist of stating a new response (i.e., sampling from background knowledge), repeating their previous partner's response, repeating their own response from the previous trial, or choosing a name from an earlier interaction (see Figure 3). In this model, background knowledge becomes progressively less important as the number of trials increases. We refer to this model as the Context Aware Agents (CAA), as it samples responses from specific content prior distributions (background knowledge about possible responses) and social context (memory trace of interaction history) based on a learned decision strategy.

Crucially, agents update their sampling strategies adaptively from both background knowledge and interaction history distributions, by following an updating procedure that weighs rewarded trials against non-rewarded trials when selecting a given decision strategy. Given the reward incentive for coordinating with networked neighborhoods, we can assess how network structure and content priors (context, hashtag versus face) impact the onset of group coherence. Our model possesses two parameters: the *learning parameter*, α, which determines how fast individuals move from using background knowledge to social context emerged from networked interaction, and the *self-preference* parameter, γ, which determines an individual's intrinsic preference for repeating themselves over their partner. See the *computational model details* section in the supplemental materials for the mathematical details of the Context Aware Agent model for more information.

Network simulation results

Simulations were generated from models fit to data gathered from the behavioral experiments, with each node in a network defined as a Context Aware Agent (see Figure 5) and interaction pairings progressing in the same manner as the experimental runs. During each trial, agents followed the decision-making pipeline illustrated in Figure SM SM 7. We compared the CAA performance with simulations from a computational model implemented by Centola and Baronchelli (2015), which served as a control model. As shown in Figure 5, the control model randomly assigns one individual from each pair to be the speaker and the other to be the hearer. If the speaker's response is already in the hearer's vocabulary, each of their vocabularies is updated to contain only that response, otherwise, the speaker's response is added to the hearer's vocabulary. The next response of each individual is a selection from their own updated vocabulary.

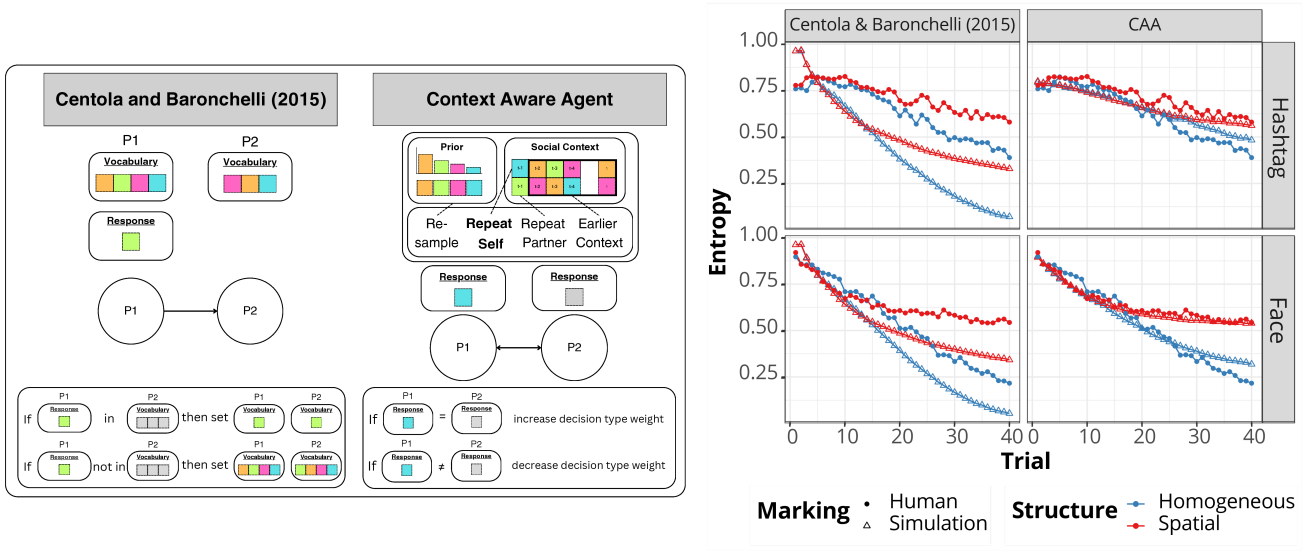

Figure 5. Overview of computational models and simulation results. **Left figure** describes the computational model from Centola and Baronchelli (2015) and the Context Aware Agent (CAA) model. The colored squares indicate possible responses in each player's vocabulary for the interaction tasks (e.g., names in the face-naming task, and hashtags in the hashtag-matching task). The illustrations show how players (P1 and P2) update their vocabulary through the interaction task. **Right figure** shows the simulation results of normalized entropy across runs for each content and structure condition against humans from both computational models.

To fit the experimental data using our CAA model, we assume that all agents utilize the same learning and self-preference parameters throughout the experiment. We optimized the learning parameter by minimizing the ℓ^2-distance on the entropy time series between the simulated and experimental results. A key observation is that participants in the hashtag experiment learn differently compared to those in the name-face experiment, yielding different learning parameter α, for these two conditions. Figure SM 8 illustrates the normalized average distance between the CAA model and experimental data as α varies, with the face experiment results represented in blue. We notice an initial decrease followed by an overall increase in distance between the CAA model's simulated results and the experimental data as α rises. Notably, there is a distinct minimum at $\alpha = 0.4$, where the second derivative is significantly greater than zero. In contrast, for the hashtag experiments, the normalized average distance decreases with α, plateauing around $\alpha = 4$. The minimum occurs at $\alpha = 4.9$, but here, the second derivative is close to zero.

The previous discussion supports the need for fitting different model parameters based on the content of interaction (hashtag vs face-naming). Figure 5 shows the entropy for all four experimental conditions. Within each panel, we present the average entropy from the human data, the control model, and the CAA model in that experimental condition. We find that the control model consistently overestimates the decrease rate in the entropy dynamics and the initial entropy. In contrast, the CAA model provides a better fit in all four experimental cases. Notably, the Context Aware Agent model matches the initial entropy observed in the experiments and performs exceptionally well for both experiments on the spatially embedded network, as shown by the red curves in the left panels. Figure 5 highlights the distance between the average experimental and model entropy values. The CAA model appears to account for human response entropy better than does the control model.

Discussion

Understanding the drivers of narrative interaction in online networks is critical for predicting how information spreads through social media and influences people's beliefs at both group and individual levels. Through a novel experimental paradigm for studying online narrative interaction and network-based computational modeling, we examined how a group's network structure interacts with individual-level decision strategies to shape the onset of shared beliefs about narrative-based evidence within a networked group. We replicated long-standing empirical findings in the face-naming task presented by Centola and Baronchelli (2015), demonstrating that information can move throughout a network in homogeneously-connected environments, facilitating shared beliefs in a group. Due to the low environmental uncertainty and low complexity of the face-naming interaction task, where less background knowledge can influence a partner's response and different responses are more interchangeable with one another (e.g., one name among many equally good ones), individuals are more likely to leverage social context to arrive at shared beliefs. Moreover, with the increase of cognitive complexity in the interaction task, different characteristics of dynamics

emerge. We found that groups explored the space of possible responses longer in a cognitively complex social interaction task (hashtag-matching), adding to the uncertainty in the social environment and limiting the onset of shared responses in homogeneously-connected networks. Extending classic findings from the communication literature on manipulating the cognitive complexity of face-to-face interaction to the online domain[23]. These findings suggest that while network-rewiring may be one route to encourage consensus in real-world networks, because of the high complexity of social media, mechanisms that encourage individuals to incorporate beliefs from network neighbors into future decisions (e.g., media engagement or content generation) will likely also be necessary to encourage narrative consensus online.

By applying a causal language model to participants' written documents, we found that network structure and narrative interaction through hashtag generation can impact how groups discuss the information conveyed in event narratives. The groups in the homogeneously-mixed hashtag-matching condition demonstrated a significant increase in mentioning a variety of causal relations from the narratives after networked interactions, suggesting that exposure to a variety of extracted information (such as hashtag) in a larger neighborhood has a potential to direct a group's attention to the range of causal content embedded in long-form narratives. Future NLP analyses of participants' personal narratives should parse a wider array of semantic relations to model shifts in situation models, the memory representations people build when processing text-based narratives[25,27,28]. This effort could elucidate how networked interactions impact people's narrative representations and illuminate mechanisms for encouraging healthier discourse online.

By fitting our agent-based model of Context-Aware Agents to decision strategy dynamics in humans, we simulated network-level dynamics for interaction tasks with both low and high levels of environmental uncertainty. However, it is important to note that we assumed all participants learn at the same rate in the experiments, which does not consider individual differences in learning. A future step in refining the data fitting is to tailor the data to individuals. This approach will not only improve the model's performance but also enable us to categorize participants into distinct clusters based on their decision strategies.

Materials and Methods

Preregistration

We preregistered the experimental design, key hypotheses, and statistical analysis framework on the Open Science Framework at the following web address https://osf.io/598dt?mode=&revisionId=&view_only=. All code and anonymized data for the presented results and software for replicating the experiments can be found at this Open Science Framework repository https://osf.io/tx6gr/ and at the following GitHub repository https://github.com/jpriniski/NetCom.

Participants

We sampled a total of $N = 1,040$ participants from the Prolific and SONA subject pools at UCLA, and placed them into one of twenty-six experimental runs. Experimental runs vary according to three factors: the *size* of a network ($N = 20, 50, 100$), its connectivity *structure* (homogeneously-mixed/fully-connected; spatially-embedded/ring-like), and the *content* of interaction (hashtag; face-name). We collected a total of twelve experimental runs for face interaction (three runs for each network structure of sizes N = 20 and N = 50), and collected fourteen experimental runs for hashtag interaction (three runs for each network structure of sizes N = 20 and N = 50, and a single run of each network structure for N = 100). Participants $N = 20$ and $N = 50$ conditions were sampled using Prolific. For the $N = 100$ condition, we recruited undergraduates in the Department of Psychology at UCLA through SONA subject pools. We posted initial recruitment surveys a week prior to each run in SONA and a few hours prior to each run in Prolific. Participants who received the most points at the end of the experiment received an additional $10 bonus.

Materials

Across all network conditions, participants first read a four-paragraph narrative description of the Fukushima nuclear disaster prior to interaction in a network. The narrative explains how a large earthquake triggered a tsunami that caused damage to a nuclear reactor and resulted in radiation leaks, population displacement, and an energy-saving movement "Setsuden". We selected this narrative based on a pilot study demonstrating that it resulted in the most diverse set of hashtags within a set of tested narratives related to natural and financial disasters. This is likely because the narrative describes a rich set of causal relations (a generative causal chain producing a branching common cause sequence) and included both negative (e.g., displacement, poisoning) and positive effects (e.g., energy saving movement). Fig. 6 illustrates the causal structure of the Fukushima disaster narrative.

Experimental Design

We used the open-source framework OTree written in Python[38], and hosted experiments on a Linux server. Participants joined the experiment through a Qualtrics survey that directed participants to the network experiment.

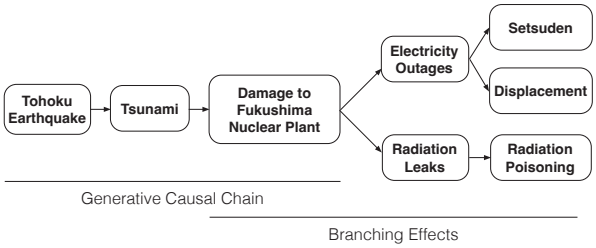

Figure 6. Causal model communicated in the nuclear disaster narrative. This diagram is just for illustration purposes, participants did not see this diagram. They read a four-paragraph narrative describing how the Tōhoku earthquake triggered a massive tidal wave that damaged the Fukushima Nuclear Power Plant, resulting in electricity outages, radiation leaks and poisoning, human displacement, and *Setsuden*, a national energy-saving holiday. The narrative is available in the OSF project website at https://osf.io/tx6gr/.

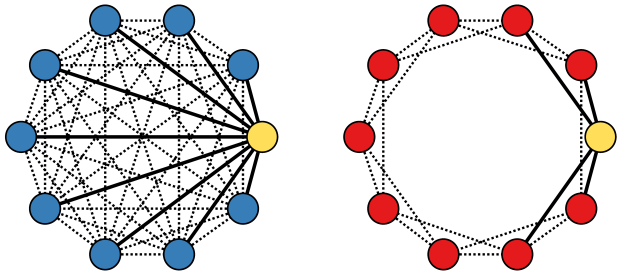

Figure 7. Two network structures tested in this experiment. Homogeneously-mixed (**left**) and spatially-embedded (**right**) networks with $N = 10$ nodes. Edges drawn with a solid line represent the neighborhoods for a hypothetical node 1 (colored yellow) in both networks. As a network's size grows, the diameter of spatial networks grow whereas homogeneous networks maintain a diameter of 1.

Our social network experiment proceeded in three steps. First, we randomly assigned each participant as a player in a network that defined who may interact with whom on a given trial. Second, we assigned interactions between individual participants on each trial. Third, we rewarded participants based on the outcome of their interactions. We can specify this process using graph theory notation. The first step is to initialize a fixed graph $G(N,E)$, defined by a set N nodes representing individual participants connected through an edge set E. We discuss below the specific graph structures used. The second step iterates over T trials. On a given trial $t \in T$, connection (edge) configurations follow mixing participants randomly within a participant's neighborhood. The third step is to identify and reward coordinated behavior. If the response from participant n_i on trial t is r_i^t, then participants n_i and n_j coordinate if $r_i^t = r_j^t$.

Procedure

The experiment consisted of three blocks: a pre-interaction block, a networked interaction block, and a post-interaction block, as shown in Figure 1. This three-block design allowed us to assess behavioral dynamics during the networked interaction block, in addition to examining whether networked interaction could shift beliefs and language outputs from the pre- to post-interaction blocks.

In the **pre-interaction block**, participants read a four-paragraph narrative describing the Fukushima nuclear disaster, and then were asked to write a "tweet" (within a 140-character limit) and ten hashtags characterizing the events described in the narrative.

In the **network interaction block**, participants joined a network experiment with real-time interaction via an online platform using the python framework OTree[38]. Participants were assigned to one of six experimental conditions based on the size of the network ($N = 20; 50; 100$) and network structure (spatially-embedded and homogeneously-mixed; see Fig. 7). Regardless of network size, nodes in spatial networks have a consistent neighborhood size $k = 4$, meaning each participant would interact with only four other participants during the entire experiment. Neighborhood size in homogeneous networks is $N - 1$, as each participant can interact with any of the remaining participants. A consequence is that the network diameter (i.e., the largest

geodesic distance in the connected network) was consistently 1 in all tested homogeneous networks, but grows as a function of size in spatial networks. Previous research showed that both features of network topology (i.e., neighborhood size and network diameter) uniquely influence the emergence of shared behavior in online networks[39].

The networked interaction block consisted of 40 trials, in which participants interacted with their partners based on the edge structure in the assigned network. On each trial, participants were instructed to write a single hashtag describing the narrative they read in the pre-interaction block. After participants submitted their hashtag response, they were then presented with a new page showing their own hashtag response, their partner's hashtag response, whether they received a point for matching responses with their partner, and their cumulative reward point. Participants were informed of their partner's response *after* submitting theirs, and were provided no additional information (see sample screenshots of an interaction trial in Figure 1). This design allowed us to measure the direct effect of coordination utilities and network structure on the production of normative network behaviors in a social network.

Following networked interactions, participants entered a **post-interaction block** in which they wrote one more "tweet" for the same narrative and another ten hashtags describing the Fukushima nuclear disaster before providing demographic information.

One consequence of these two network structures is not only who is connected to who, but also the amount of repeated interactions a participant has with their neighborhood. The number of times a participant interacts with their full set of neighbors a total of $\frac{T}{N-1}$ times before completing the experiment. This means, in the fully-connected condition (i.e., homogeneouly-mixed), participants interacted with their full set of neighborhood 2 times when $N = 20$, 81.6% of their neighbors when N = 50, and 40.4% of their neighbors when N = 100. Meanwhile, participants interacted with their fully set of neighbors 10 times in the spatially-embedded structures. These conditions allow us to contrast the effect of neighborhood size relative to ties across the network, and to determine the impact of repeated interactions between pairs of partners to produce dominant behaviors (e.g., participants in the network responding in a consistent manner).

References

1. Sasahara, K. *et al.* Social influence and unfollowing accelerate the emergence of echo chambers. *J. Comput. Soc. Sci.* **4**, 381–402 (2021).
2. Priniski, J. H., McClay, M. & Holyoak, K. J. Rise of qanon: A mental model of good and evil stews in an echochamber. *arXiv preprint arXiv:2105.04632* (2021).
3. Centola, D. The network science of collective intelligence. *Trends Cogn. Sci.* **26**, 923–941 (2022). Publisher: Elsevier.
4. Centola, D. & Baronchelli, A. The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proc. Natl. Acad. Sci.* **112**, 1989–1994 (2015).
5. Centola, D. The Spread of Behavior in an Online Social Network Experiment. *Science* **329**, 1194–1197 (2010). Publisher: American Association for the Advancement of Science.
6. Priniski, J. & Holyoak, K. J. A darkening spring: How preexisting distrust shaped COVID-19 skepticism. **17**, e0263191, DOI: 10.1371/journal.pone.0263191 (2022). Publisher: Public Library of Science.
7. Priniski, J. H. *et al.* Mapping moral valence of tweets following the killing of george floyd. *arXiv preprint arXiv:2104.09578* (2021).
8. Bennett, W. L. & Livingston, S. Platforms, politics, and the crisis of democracy: Connective action and the rise of illiberalism. *Perspectives on Polit.* 1–20 (2025).
9. Papacharissi, Z. *Affective publics: sentiment, technology, and politics*. Oxford studies in digital politics (Oxford University Press, New York, NY, 2015).
10. Tangherlini, T. R., Shahsavari, S., Shahbazi, B., Ebrahimzadeh, E. & Roychowdhury, V. An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web. *PLOS ONE* **15**, e0233879 (2020).
11. Giaxoglou, K. #jesuischarlie? hashtags as narrative resources in contexts of ecstatic sharing. *Discourse, context & media* **22**, 13–20 (2018).
12. Dawson, P. Hashtag narrative: Emergent storytelling and affective publics in the digital age. *Int. J. Cult. Stud.* **23**, 968–983 (2020).
13. Yang, G. Narrative Agency in Hashtag Activism: The Case of #BlackLivesMatter. *Media Commun.* **4**, 13–17 (2016).
14. Papacharissi, Z. Affective publics and structures of storytelling: sentiment, events and mediality. *Information, Commun. & Soc.* **19**, 307–324 (2016).

15. Howard, P. N. & Hussain, M. M. *Democracy's Fourth Wave?: Digital Media and the Arab Spring* (Oxford University Press, 2013).

16. Zappavigna, M. Searchable talk: The linguistic functions of hashtags. *Soc. semiotics* **25**, 274–291 (2015).

17. Booten, K. Hashtag drift: Tracing the evolving uses of political hashtags over time. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2401–2405 (ACM, 2016).

18. Lin, Y.-R., Margolin, D., Keegan, B., Baronchelli, A. & Lazer, D. # bigbirds never die: Understanding social dynamics of emergent hashtags. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, 370–379 (2013).

19. Cunha, E. *et al.* Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proceedings of the workshop on language in social media (LSM 2011)*, 58–65 (2011).

20. Avolio, M. L. *et al.* A comprehensive approach to analyzing community dynamics using rank abundance curves. *Ecosphere* **10**, e02881 (2019).

21. Hallett, L. M. *et al.* codyn: An r package of community dynamics metrics. *Methods Ecol. Evol.* **7**, 1146–1151 (2016).

22. Barkoczi, D. & Galesic, M. Social learning strategies modify the effect of network structure on group performance. *Nat. communications* **7**, 13109 (2016).

23. Delia, J. G., Clark, R. A. & Switzer, D. E. Cognitive complexity and impression formation in informal social interaction. *Commun. Monogr.* **41**, 299–308 (1974).

24. Bieri, J. Cognitive complexity-simplicity and predictive behavior. *The J. Abnorm. Soc. Psychol.* **51**, 263–268, DOI: 10.1037/h0043308 (1955).

25. Morrow, D. G., Bower, G. H. & Greenspan, S. L. Updating situation models during narrative comprehension. *J. memory language* **28**, 292–312 (1989).

26. Zwaan, R. A., Magliano, J. P. & Graesser, A. C. Dimensions of situation model construction in narrative comprehension. *J. experimental psychology: Learn. memory, cognition* **21**, 386 (1995).

27. Zwaan, R. A., Langston, M. C. & Graesser, A. C. The construction of situation models in narrative comprehension: An event-indexing model. *Psychol. science* **6**, 292–297 (1995).

28. Zwaan, R. A. & Radvansky, G. A. Situation models in language comprehension and memory. *Psychol. bulletin* **123**, 162 (1998).

29. Priniski, J., Verma, I. & Morstatter, F. Pipeline for modeling causal beliefs from natural language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 436–443 (Association for Computational Linguistics, Toronto, Canada, 2023).

30. Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794* (2022).

31. Liu, Y. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* **364** (2019).

32. Fudenberg, D. & Tirole, J. *Game Theory* (MIT Press, Cambridge, MA, 1991).

33. Neumann, J. V. & Morgenstern, O. *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, NJ, 1944).

34. Taylor, P. D. & Jonker, L. Evolutionary stable strategies and game dynamics. *Math. Biosci.* **40**, 145–156, DOI: 10.1016/0025-5564(78)90077-9 (1978).

35. Klein, G. Naturalistic decision making. *Hum. Factors* **50**, 456–460, DOI: 10.1518/001872008X288385 (2008).

36. Tversky, A. & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science* **185**, 1124–1131, DOI: 10.1126/science.185.4157.1124 (1974).

37. Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Annu. Rev. Psychol.* **62**, 451–482, DOI: 10.1146/annurev-psych-120709-145346 (2011).

38. Chen, D. L., Schonger, M. & Wickens, C. oTree—An open-source platform for laboratory, online, and field experiments. *J. Behav. Exp. Finance* **9**, 88–97 (2016).

39. Anagnostopoulos, A., Becchetti, L., Castillo, C., Gionis, A. & Leonardi, S. Online team formation in social networks. In *Proceedings of the 21st international conference on World Wide Web*, 839–848 (2012).

40. McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (Chapman and Hall/CRC, New York, 2016).

41. Heiss, A. A guide to modeling outcomes that have lots of zeros with Bayesian hurdle lognormal and hurdle Gaussian regression models, DOI: 10.59350/ety2j-09566.

Acknowledgments

This work was funded in part by the AFOSR MURI grant No. FA9550-22-1-0380 and the DARPA Army Research Office (ARO), under Contract No. W911NF-21-C-0002. Rodriguez was partially funded by NSF-DMS-2042413. We thank Yiling Yun, Yiting Wang, and Chloe Ji for their help in running the experiments; Caleb Musfeldt and Godwina Ogbeide for their comments on the manuscript; and Zachary Kilpatrick for his helpful discussions.

Author contributions statement

HP formulated psychological theory, advised the development of network experiment software, performed statistical analyses on behavioral data, administered the experiments, wrote draft and edited manuscript. BL administered experiments and recruited participants. AH formulated, implemented, and described the computational model. NR advised computational modeling, formulated modeling theory, edited the manuscript, and secured funding. SK implemented network experiment software, and FM advised the development of network experiment software. JB advised on theory, analyses, grant funding. HJ formulated the psychology theory, advised on software development, experimental design, analyses of behavioral and simulation data, secured funding.

Supplemental Information

Computing normative responses

In Figure 2 from the main text above, we display the onset of shared response in each experimental condition. Shared normative responses, represent the agreement/consensus of a group. Following previous literature[4], we encode a normative response by dividing the number of respondents in a group who produced the most popular response on a given trial by the group size. The response could change from trial to trial.

Analytic framework for statistical modeling of normative response

We fit Bayesian generalized linear models (GLM) to predict how the two network structures (spatially-embedded vs homogeneously-mixed structure) and the content of interaction (hashtag response vs face-naming) support the emergence of group coherence. We fit separate models to predict the proportion of a group producing a dominant response, and the entropy of the group's full response distribution. We assume that the proportion of participants who produced the dominant hashtag on trial t follows a Beta distribution, a commonly used distribution to predict proportion data[40]; we used uninformative priors (i.e., $\mathcal{N}(0, 10)$) for regression coefficients. Specifically, the GLM model predicted the coherence value as a function of trial number (i.e., *Trial*) interacted with network structure (*Spatial* vs *Homogeneous*) and content (*Hashtag* vs *Face-name*), while controlling for network *size*.

Prior distribution of responses illuminates background knowledge about interaction content

As shown in Figure SM 1, the prior for face names is more distributed than that for hashtags, which shows that responses consolidated around broad topic labels (e.g., #NuclearDisaster) and causal/generative events (e.g., #Earthquake and #Tsunami) on the first trial in networked interaction. The prior distribution of possible responses influences how participants choose to explore background knowledge within a social context (i.e., responses sampled from prior interactions that yielded rewards) when coordinating with network neighbors. The high uncertainty in the distribution of face names promotes more reliance on learning from social context during networked interactions, which enables a quicker onset of group-level consensus.

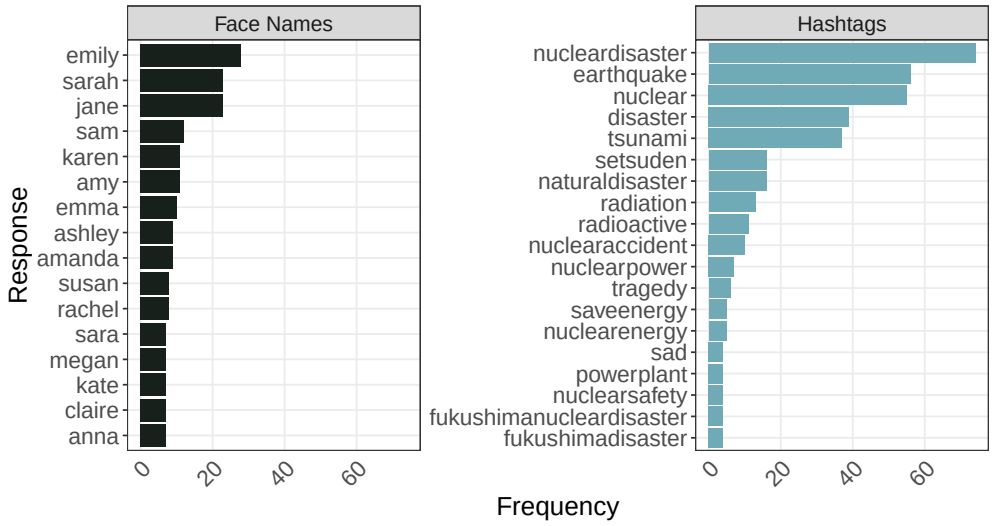

Figure SM 1. Distribution of the top 15 most common responses in the first trial of face naming and hashtag matching experimental runs. In the first trial, face name responses exhibit a widely dispersed distribution, while hashtag responses show a more skewed distribution with increased frequency of hashtags that describe broad topics (e.g., Nuclear Disaster) and causal topics expressed in the disaster narrative. The increased skewness in hashtag interactions suggests that participants use information about the event extracted from the narrative when starting interaction on this task. In addition to background knowledge, there are multiple strings mapping onto the same narrative entities (e.g., the several hashtags related to the nuclear disaster topic). These two factors make it more difficult for groups to align on shared responses.

Entropy dynamics of response distribution

Entropy encodes the overall coherence of a response distribution. In Figure SM 2, we observe that the entropy approaches zero more quickly under the face-naming condition. This suggests a faster shift toward coherence across responses compared to the hashtag condition.

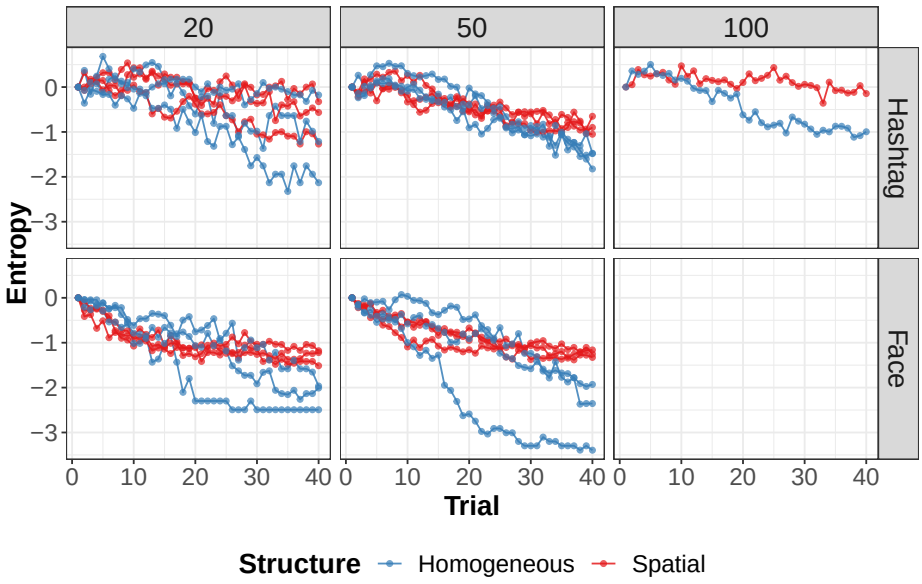

Figure SM 2. Entropy change of response distribution over course of interaction by group size (columns) and interaction content (rows). The entropy values represent the change in entropy in trial t from the first trial, allowing normalization between group sizes (larger group sizes inherently have higher entropy due to larger distribution of responses).

Cognitive complexity shapes onset of local coordination between interacting nodes

The group-level coherence findings described above replicated well-known findings in behavioral economics and sociology suggesting that homogeneously-mixed networks better produce shared behaviors than spatially-embedded networks. We not only replicated previous findings using materials with low cognitive complexity in our face-naming condition[4], but extended them to interactions with high cognitive complexity in the matching hashtag condition. However, even when a spatially-embedded group does not adopt a shared behavior, participants could be aligning responses with their neighbors. Forming disjoint clusters akin to echo chambers with separable social groups aligning on different behaviors (e.g., labeling a narrative with opposing causal content).

As shown in Figure SM 3, there is a steady increase in coordination rate (defined in the main text) in each of the conditions, suggesting that subjects learn to match responses in both conditions. To test how the cognitive complexity of interaction affects the onset of local coordination, we fit a Bernoulli GLM to predict the probability a pair of nodes (i.e., partnered participates in a trial) coordinate responses by interacting trial number with network structure and interaction content, while controlling for group size. As shown in Figure SM 3, at the beginning of an experiment, participants in the reference group (homogeneously-mixed face-naming networks of size 20) coordinate on approximately 5% of trials ($\beta_{Intercept} = -2.83$, 95% CI $[-2.98, -2.70]$), with a coordination rate increasing by approximately 0.5% for each subsequent interaction ($\beta_{Trial} = 0.10$, 95% CI $[0.09, 0.10]$). Although participants in spatially-embedded networks coordinated more effectively than those in homogeneously-mixed networks ($\beta_{Spatial} = 1.46$, 95% CI $[1.29, 1.63]$), the adoption of shared behaviors in homogeneously-mixed networks is accompanied by faster learning to coordinate ($\beta_{Trial:Spatial} = -0.05$, 95% CI $[-0.05, -0.04]$). The cognitive complexity of the interactions affected coordination rates, as groups learned to coordinate face names more quickly than hashtags ($\beta_{Trial:Hashtag} = -0.04$, 95% CI $[-0.05, -0.04]$). The cognitive complexity of the interactions also mediated the impact of network structure on coordination dynamics ($\beta_{Trial:Hashtag:Spatial} = 0.04$, 95% CI $[0.03, 0.05]$).

Statistical models for decision strategy dynamics

To predict the number of participants following one of the four decision strategies in a given network structure and content interaction condition, we fit a Bayesian GLM with a categorical response distribution. The analysis revealed that participants in hashtag-matching networks were more likely to explore new responses than follow decision strategies that exploit social context.

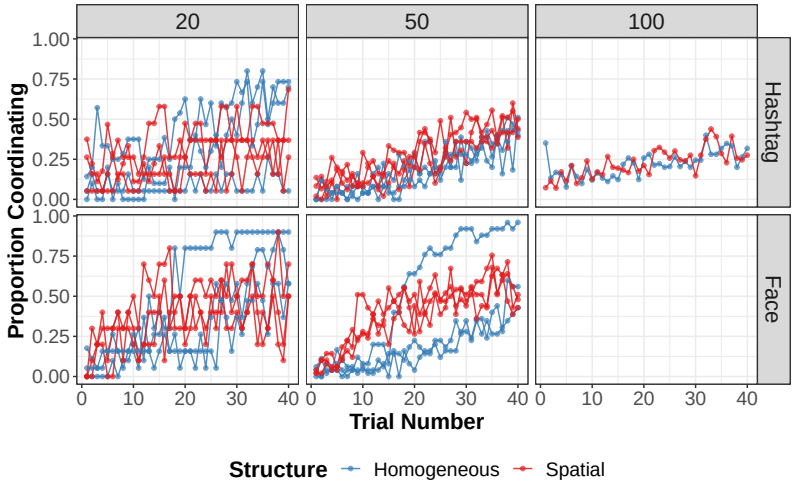

Figure SM 3. Proportion of nodes in a group coordinating on each trial.

These models are fit to the proportion of participants sampling each of the four strategies, displayed in Figure 3. Participants in hashtag matching networks were less likely to repeat themselves (RS) ($\beta_{RS:Hashtag} = -1.38$, 95% CI $[-1.49, -1.26]$) and partners (RP) ($\beta_{RP:Hashtag} = -0.98$, 95% CI $[-1.10, -0.85]$), as well as sample a response remembered from earlier context (EC) ($\beta_{EC:Hashtag} = -0.40$, 95% CI $[-0.55, -0.26]$).

Impact of cognitive complexity on response coordination

Prominent game theory models used to study the emergence of coordinated behavior in networks typically account for the role utility representation plays in taking certain actions over others, but do not consider the impact of representational content connected to those utilities. Effective communication requires semantic and causal/categorical alignment across group members to facilitate social learning (e.g., converging on a shared language-response, social norm). As shown in Figure SM 4, the top hashtags for coordination are a mix of semantic topics and cause-effect relational entities encoded by the narratives situation model. Semantic and causal content constrain the response space and guide what constitutes an optimal response. Moreover, studies of naturally occurring hashtag behavior revealed that successful hashtags fall into one of two categories: focal hashtags, which tag posts with broad semantic topics to relate them to larger discussions and movements across an online network (e.g., MeToo, BlackLivesMatter, SeoulCrowdCrush), and individualistic hashtags, which make the distribution of hashtags heavy-tailed, as they co-occur with focal hashtags while allowing users to signal personal narratives (e.g., MeTooSurvivor, BLMProtest, PrayForSeoul)[17]. In our experiment, the top three most successful hashtags align on the same discrete narrative entity, but have different string representations. Not only does causal complexity of the hashtag content make it more difficult to coordinate responses and raises the cognitive complexity of interaction, but also the role that additional appending of semantic content to signal a personal view (representation of narratives) has on hashtag coordination.

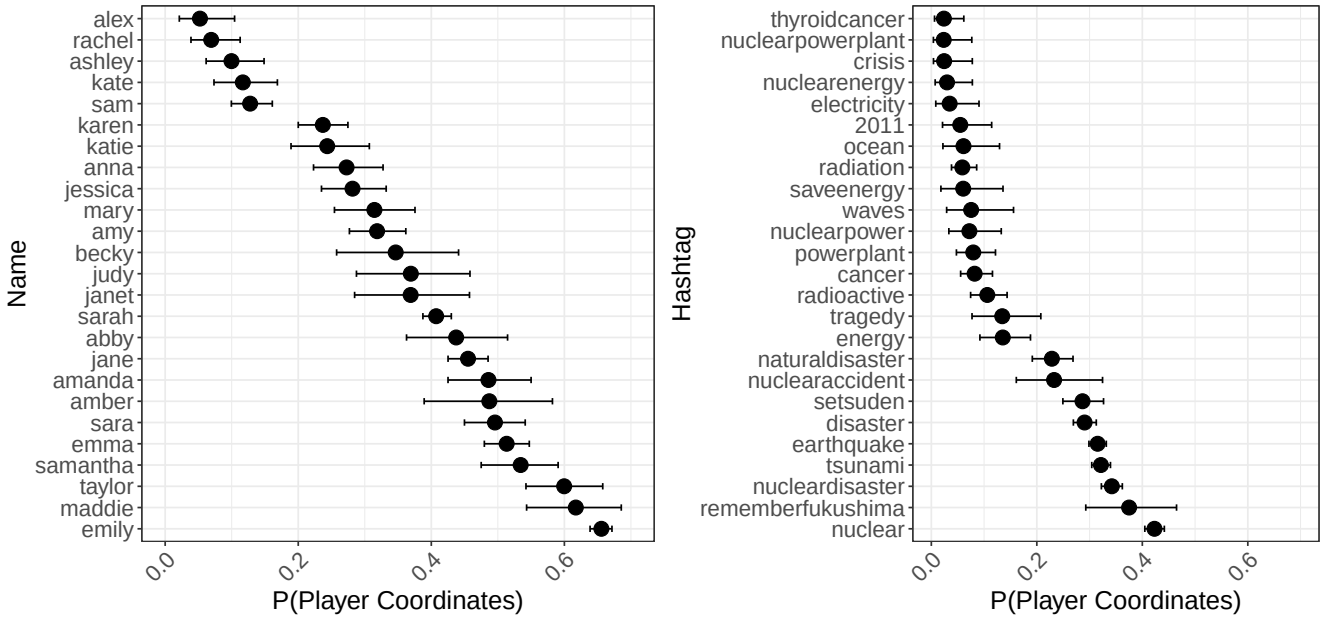

Figure SM 4. Probability distribution of player coordination **Left:** Probability of a participant coordinating given a sampled face. The analysis focuses on trials that contained one of the top 25 faces.**Right:** Probability of a participant coordinating given a sampled hashtag. The analysis focuses on trials that contained one of the top 25 hashtags.

Modeling causal language identified in personal narratives

The causal language analysis pipeline identifies causal tuples in each topic. A document is labeled as having a causal relation if there is a span of tokens belonging to a cause and a span of tokens belonging to an effect within the document. The algorithm then finds causal topics by clustering the cause-and-effect spans based on their semantic topics (see Supplemental Table SM 1). Therefore, extracted topics tend to belong more to cause events or more to effect events. We show the distribution of causes and effects for each of the clusters in Figure SM 5. In our main analysis on causal language change, we examine how these topic distributions change in personal narratives composed from pre- to post-interaction in our experimental networked environment.

Causal language change in personal narratives following network interaction

As shown in Figure SM 6, and discussed in the main part of the manuscript, the distributions of the shifts in the amount of causal language for each participant (the number of causal claims generated after interaction subtracted by the number of causal claims generated before interaction) are zero-inflated. The red line on each of the plots indicates the mean of each distribution. The negative shift values in the spatial hashtag condition explain why the t-statistic is lower than the others in Figure 4. Despite these negative values (which suggest that more participants provided less causal claims after interaction than before interaction), does not overtake the central shift towards the initial causal chain in the narrative after the networked interaction, which is shown in the causal relation level diagram in the main text of the manuscript.

We modeled the distribution of shifts shown in 4 in causal claims following networked interaction by fitting fit a hurdle Gaussian model to predict the shift in the number of causal claims that a participant produced after network interaction (difference = number of causal claims produced after interaction - number of causal claims produced before interaction). The hurdle Gaussian model consists of a logistic classification step to identify personal narratives without causal claims, and then a Gaussian distribution estimating the difference scores for the remaining documents[41]. We examined how network structure and content in networked interaction impact the difference scores.

The hurdle Gaussian model reveals that around 49% of the participants did not show a change in the number of causal claims in personal narratives after networked interaction ($hu = .49$, 95% CI [0.46, 0.52]). The intercept of the Gaussian linear model component equals the mean change in the number of causal claims for participants placed in homogeneously-mixed face-naming networks, which is not credibly different from zero ($\beta_0 = .21$, 95% CI [-0.10, 0.54]). In line with our prediction that interaction content will have a main effect on the generation of causal content in personal narratives, we found a significant effect of hashtag interaction on change in causal language ($\beta_{Hashtag} = .45$, 95% CI [0.00, 0.88]).

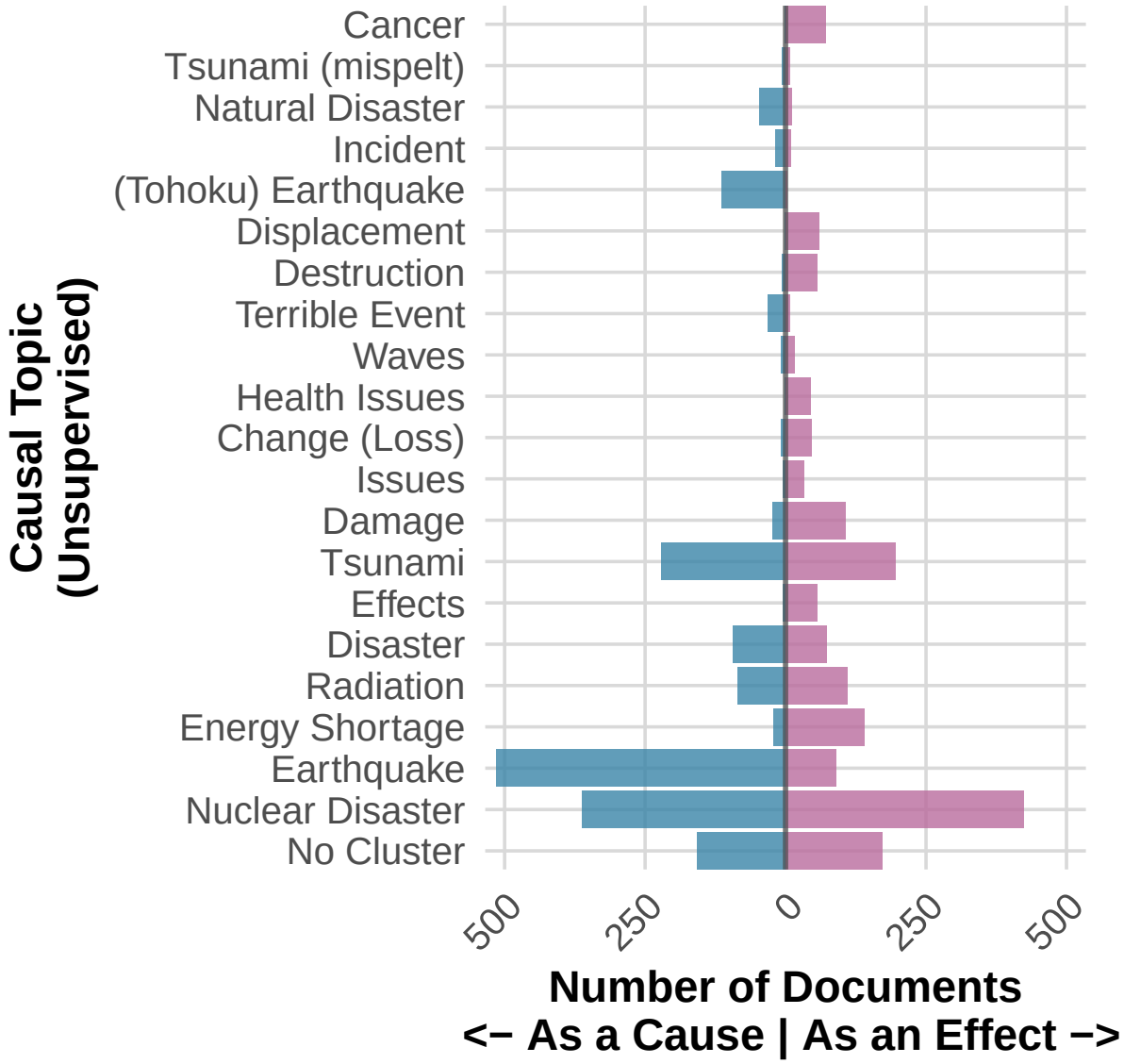

Figure SM 5. Distribution of documents instantiating each causal topic as a cause and effect in causal relations identified by the causal language model. Causal topic labels are the result of an unsupervised clustering algorithm as descried in Priniski et al., (2024). Causal topic labels are listed on the y-axis, and the number of documents mentioning that topic as a cause (left-shooting blue bars) and effect (right-shooting purple bars). The unsupervised clustering algorithm returns multiple topics that we reduced to map onto the causal model of the Fukushima narrative (e.g., mapping Tohoku Earthquake and Earthquake onto the same label). See Table SM 1 for more details about causal topic remapping, cluster alignment with the narrative materials, and keywords for each topic.

Significantly shifted causal relations following network interaction

As discussed in the main text, we performed a t-test to identify significantly shifted causal relationships expressed in personal narratives following network interactions. Table SM 2 shows the full list of significantly shifted edges. Figure 4 shows positive shifts for clarity, but these positive shifts can result from a shift away from other relationships or from self-referential clusters, both of which are not shown in the main manuscript. The complete list of significant values for each of the 20 $\times$ 20 topics can

Remapped	Narrative Entity	Unsupervised	ID	Keywords
Earthquake	Yes	(Tohoku) Earthquake	15	tohoku earthquake, the 2011 tohoku earthquake
		Earthquake	1	massive earthquake, the earthquake
Tsunami	Yes	Tsunami	6	tsunami, a tsunami, the tsunami
		Tsunami (misspelt)	18	a large tsunami, a tsunami
		Waves	11	a tidal wave, 130 foot waves, 130 foot tsunami
Nuclear Disaster	Yes	Nuclear Disaster	0	nuclear disaster, fukushima nuclear disaster
		Damage	7	the damage, intense damage, widespread damage
Electricity Outage	Yes	Change (Loss)	9	reducing, loss, outage
Radiation Leaks	Yes	Radiation	3	radiation, radioactive isotopes, particles
Setsuden	Yes	Energy Movement	2	Setsuden, energy crisis, conserving electricity
Poisoning	Yes	Health Issues	10	health concerns, health problems, many illnesses
		Cancer	19	thyroid cancer, cancer, thyroid issues
Displacement	Yes	Displacement	14	displaced, the displacement, displaced people
Disaster	No	Disaster	4	a disaster, the disaster, many disasters
Effects	No	Effects	5	harm, environmental damage, devastating effects
Issues	No	Issues	8	problems, issues, fails, many problems
Terrible Event	No	Terrible Event	12	the event, terrible events, cataclysmic events
Destruction	No	Destruction	13	destruction, destroyed, the devastation
Incident	No	Incident	16	an accident, devastating accident, an incident
Natural Disaster	No	Natural Disaster	17	a natural disaster, this natural disaster

Table SM 1. Causal topics identified by the causal language model. The causal topic model identified 20 topics plus a catch-all "no cluster" topic (not shown). Each of the topics aligned with at least one of the narrative entities described in the Fukushima disaster materials, along with additional semantic topics. In some cases, the model identified separate clusters that aligned with the same narrative entity described in the materials (unsupervised topics), which we mapped to the underlying narrative (remapped). Statistical analyses were performed on the remapped topics. The keywords represent the most common entities in each of the unsupervised topics.

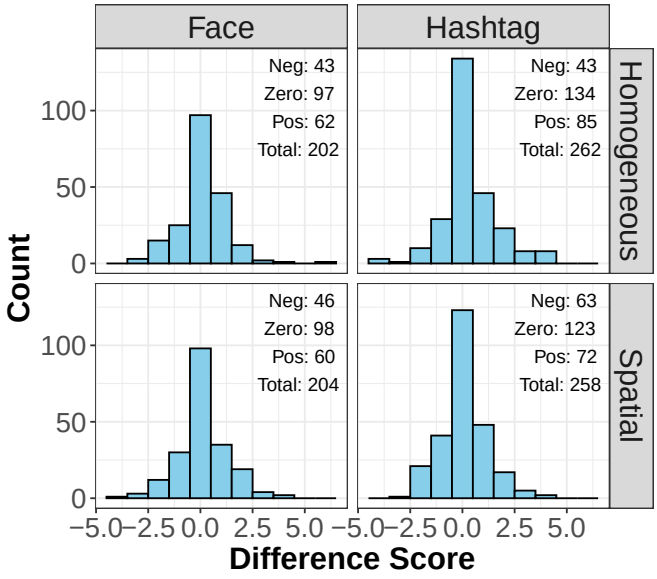

Figure SM 6. Distribution of changes in the amount of causal language produced in each experimental condition. The text in the top right of each panel describes the number of documents exhibiting a negative shift (more causal claims before network interaction compared to after), zero shift (the same amount of identified causal relations before and after), and positive shift (more causal relations after network interaction than before). For each condition, the shift is zero-inflated, indicating that many participants did not exhibit a shift in causal language.

be found at this project's repository on the Open Science Framework.

Structure	Content	Cause Topic	Effect Topic	Estimate	p	Conf Int
Homogeneous	Face	Health Issues	Health Issues	0.0584	0.0109	(0.0137, 0.1031)
Homogeneous	Face	No Cluster	Energy Shortage	-0.0292	0.0451	(-0.0577, -0.0006)
Homogeneous	Hashtag	Earthquake	Radiation	0.0299	0.0249	(0.0038, 0.0561)
Homogeneous	Hashtag	Earthquake	Tsunami	0.0898	0.0050	(0.0275, 0.1522)
Homogeneous	Hashtag	Natural Disaster	Nuclear Disaster	0.0240	0.0452	(0.0005, 0.0474)
Homogeneous	Hashtag	Nuclear Disaster	Nuclear Disaster	-0.0599	0.0180	(-0.1093, -0.0104)
Homogeneous	Hashtag	Tsunami	Destruction	0.0240	0.0452	(0.0005, 0.0474)
Homogeneous	Hashtag	Tsunami	Nuclear Disaster	0.0958	0.0015	(0.0372, 0.1545)
Homogeneous	Hashtag	Tsunami	Radiation	0.0240	0.0452	(0.0005, 0.0474)
Spatial	Face	Earthquake	Tsunami	0.0863	0.0068	(0.0242, 0.1485)
Spatial	Face	Energy Shortage	Energy Shortage	-0.0432	0.0334	(-0.0829, -0.0034)
Spatial	Hashtag	Earthquake	Tsunami	0.0674	0.0139	(0.0139, 0.1210)
Spatial	Hashtag	Energy Shortage	Energy Shortage	-0.0393	0.0192	(-0.0722, -0.0065)
Spatial	Hashtag	Tsunami	Nuclear Disaster	0.0730	0.0089	(0.0185, 0.1276)

Table SM 2. Results of t-tests for significantly shifted causal relations in each of the experimental conditions. The estimate and confidence interval columns indicates the estimated mean shift for that specific causal relation following network interaction, and p values are for tests that the mean shift is zero.

Additional information for the computational network modeling of group behavior

Agent response strategy updating and interaction patterns

As shown in Figure (SM), our model is composed of a network of Context Aware Agents (CAA), where context priors are fit to experimental data (i.e., responses across all experimental first trials), and where an individual agent's probabilities of sampling a specific strategy on a given trial is calculated as follows:

$$P(\text{BN}) = \frac{\alpha}{\alpha + T + C}, \quad P(\text{EC}) = (1 - P(\text{BN})) \cdot \frac{p_{ec} + 3}{p_{ec} + p_{rs} + p_{rp} + 20},$$
$$P(\text{RP}) = (1 - P(\text{BN}) - P(\text{EC})) \cdot \gamma \cdot \frac{p_{rp} + 1}{p_{rp} + p_{rs} + 2}, \quad P(\text{RS}) = 1 - P(\text{BN}) - P(\text{EC}) - P(\text{RP}),$$

where T is the current time-step, $\alpha \in (0, \infty)$ is the so-called "learning parameter" which dictates the speed of the network to move away from using the "brand new" strategy, C is the number of points scored per individual weighted by the time those points were scored, p_{ec}, p_{rs}, and p_{rp} are the time-weighted points using earlier context, repeat self, and repeat partner decision strategies, respectively, and $\gamma \in [0, 1]$ is the self-valuation parameter, conferring a possible proclivity to favor repeating yourself over repeating your partner. A diagram illustrating the structure of this decision pipeline is shown in SM 7.

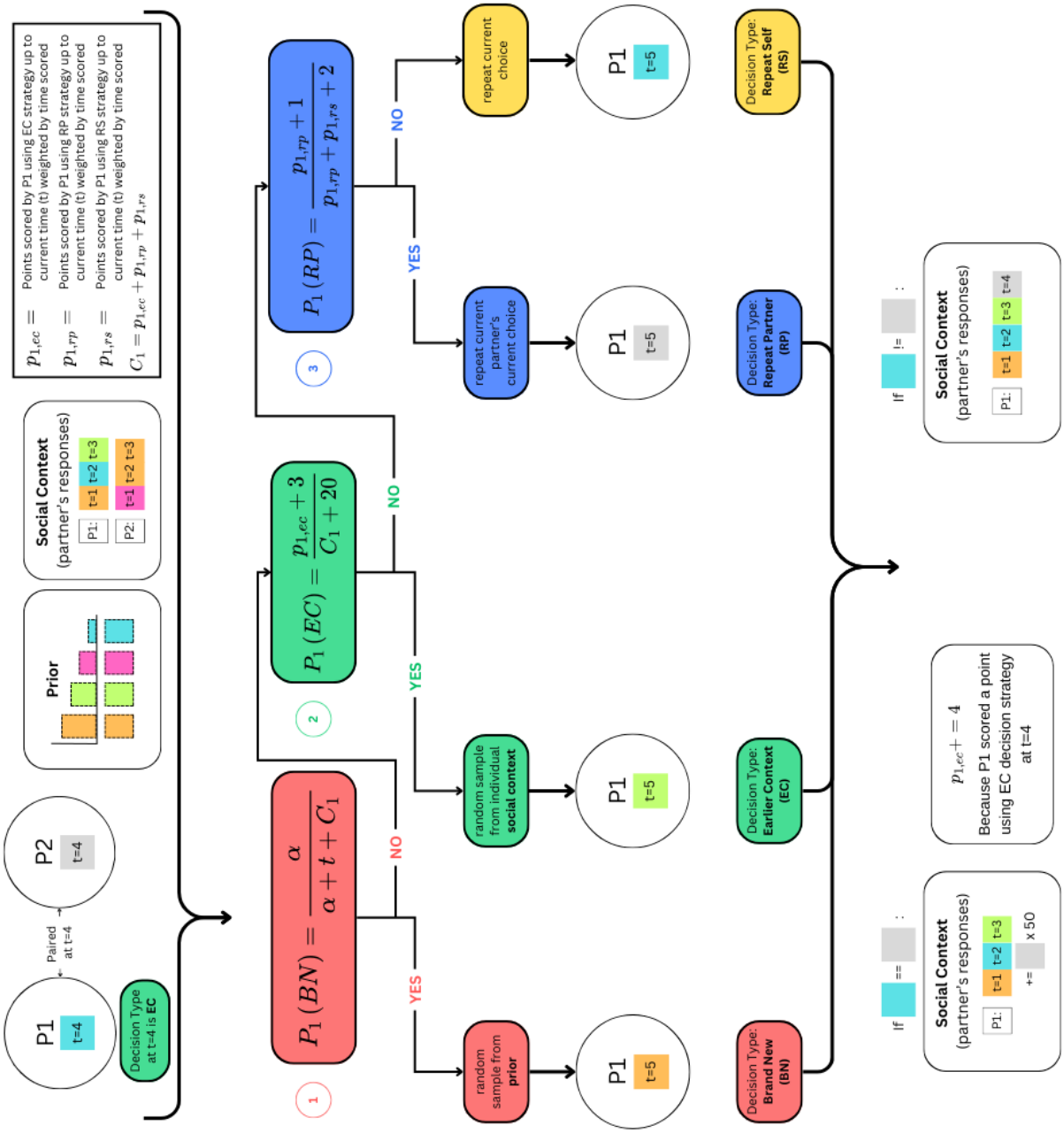

Figure SM 7. Illustration of Decision Making Model. This diagram illustrates the decision-making flow for P1 at time t=5 given the input information. Agent decisions are independent. While probabilities are necessarily dependent on the *responses* of other agents in the network, they are not at all dependent on the *decision strategies* of the other individuals. This is a necessary component of the model to facilitate appropriate parameter optimization.

Pseudocode for computational models

The pseudocode for the two computational decision-making models tested in this paper is shown below. The code for both models is on the project's repository on the Open Science Framework (see Preregistration section in the main text).

Algorithm 1 Centola Decision-Making Algorithm

for each round t in number of rounds **do**
  for each pair $(\text{agent1}, \text{agent2})$ in pairings **do**
    Randomly assign speaker and hearer roles to agent1 and agent2
    if speaker's current response is in hearer's vocabulary **then**
      speaker's vocabulary $\leftarrow$ {speaker's response}
      hearer's vocabulary $\leftarrow$ {speaker's response}
    else
      Add speaker's response to hearer's vocabulary
      agent1's new response $\leftarrow$ random selection from agent1's vocabulary
      agent2's new response $\leftarrow$ random selection from agent2's vocabulary

Algorithm 2 CAA Decision-Making Algorithm

for each round t in number of rounds **do**
  for each pair $(\text{agent1}, \text{agent2})$ in pairings **do**
    if agent1's response = agent2's response **then**
      Update each agent's scores (and relevant decision-type scores)
      for each agent **do**
        decision type is BN with probability $\frac{\alpha}{\alpha + t + \text{time-weighted context points}}$ ▷ time-weighted context points = $\sum_t$ context sampling resulted in point
      if decision type is BN **then**
        agent's new response $\leftarrow$ random sample from experiment prior
      else
        decision type is EC with probability $\frac{\text{EC points} + 3}{\text{time-weighted context points} + 20}$
        if decision type is EC **then**
          agent's new response $\leftarrow$ random sample from agent's context
        else
          decision type is RS with probability $\frac{\text{RS points} + 1}{\text{RS points} + \text{RP points} + 2}$
          if decision type is RS **then**
            agent's new choice $\leftarrow$ agent's current choice
          else ▷ decision type is RP
            agent's new choice $\leftarrow$ partner's current choice
            if agent1's response = agent2's response **then**
              for each agent **do**
                Add $50 \times t$ instances of partner's response to their context
                ▷ agents favor responses that have scored them points

              for each agent **do**
                Add t instances of partner's response to their context

Optimizing model hyper-parameters with human data

Participants adopted different response strategies in different interaction content conditions which impacted the onset of group-level consensus (i.e., participants were more likely to leverage previously encountered and generated responses in the face naming condition than in the hashtag matching condition). Our computational models included an α parameter that constrained how often an agent samples new responses versus re-samples previous responses over the course of network interactions. As shown in Figure SM 8, we plot the average distance between simulated and experimental entropy vectors for each of the four experiment types while controlling for network size. We see that the best-fitting exploration parameter for face-naming networks ($\alpha = 0.40$), is much lower than the best-fitting exploration parameter for hashtag-matching networks ($\alpha = 4.90$). Furthermore, for these alpha parameters, there is no dramatic difference across network structures (up versus down pointing arrows). These results suggest that interaction content is the salient factor driving exploration strategies in human groups, and lends support to the cognitive complexity hypothesis that individuals are slower to exploit environmental regularities in complex interaction tasks which limits the onset of group-level consensus.

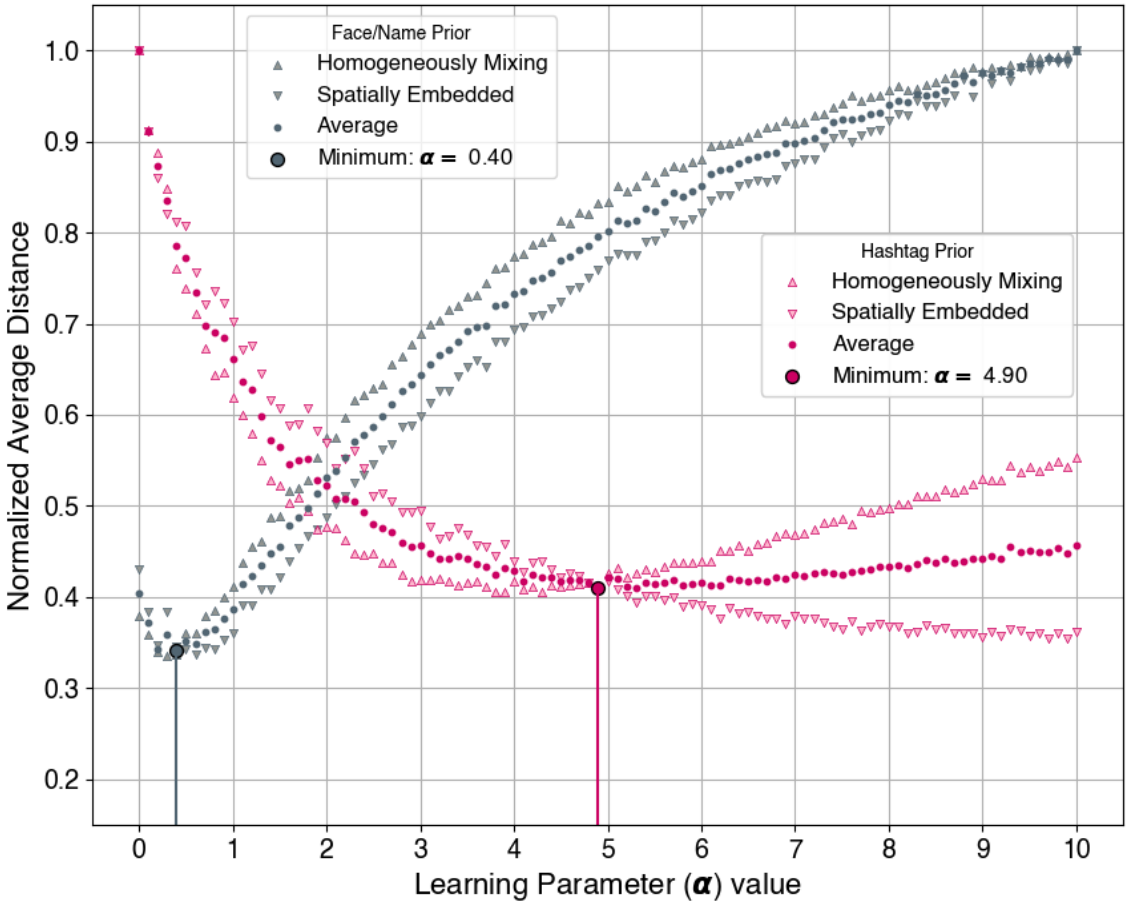

Figure SM 8. Normalized average distance in the group level response entropy between the CAA model and experimental runs conditioned on a specified value of α, which constrains how much agents re-sample previously encountered and generated responses on a present trial. Model fits for face-naming groups is illustrated in gray, while model fits for hashtag-matching groups is in pink.