

Realistic threat perception drives intergroup conflict: A causal, dynamic analysis using generative-agent simulations

Suhaib Abdurahman^{1,2*}, Farzan Karimi-Malekabadi^{1,2},
Chenxiao Yu³, Nour S. Kteily⁴, Morteza Dehghani^{1,2,3}

¹*Department of Psychology, University of Southern California.

²Center for Computational Language Sciences, University of Southern California.

³Department of Computer Science, University of Southern California.

⁴Kellogg School of Management, Northwestern University.

*Corresponding author(s). E-mail(s): sabdurah@usc.edu;
Contributing authors: karimima@usc.edu; cyu96374@usc.edu;
n-kteily@kellogg.northwestern.edu; mdehghan@usc.edu;

Abstract

Human conflict is often attributed to threats against material conditions and symbolic values, yet it remains unclear how they interact and which dominates. Progress is limited by weak causal control, ethical constraints, and scarce temporal data. We address these barriers using simulations of large language model (LLM)-driven agents in virtual societies, independently varying realistic and symbolic threat while tracking actions, language, and attitudes. Representational analyses show that the underlying LLM encodes realistic threat, symbolic threat, and hostility as distinct internal states, that our manipulations map onto them, and that steering these states causally shifts behavior. Our simulations provide a causal account of threat-driven conflict over time: realistic threat directly increases hostility, whereas symbolic threat effects are weaker, fully mediated by ingroup bias, and increase hostility only when realistic threat is absent. Non-hostile intergroup contact buffers escalation, and structural asymmetries concentrate hostility among majority groups.

Keywords: intergroup conflict, threat perception, generative agents, social simulation

1 Introduction

Human history is marked by groups fighting over material concerns such as resources or physical security, and over symbolic concerns such as identity and sacred values. In psychology, this distinction is formalized by Integrated Threat Theory [ITT; 1] as a contrast between *realistic threats* and *symbolic threats*. Political science frames the same divide in realist versus constructivist theories of power and identity and in the “greed versus grievance” debate in civil war studies [2–6]. Sociology and history similarly contrast material opportunity structures with symbolic boundaries and cultural frames [7, 8], and work in neuroscience suggests partially distinct processing for threats to sacred values relative to material threats [9]. Yet, despite decades of research across these disciplines, it remains unclear how realistic and symbolic threats drive group conflict, whether one tends to dominate, whether they amplify or subsume one another. Clarifying this relationship would help reconcile competing theories of conflict and improve our ability to understand, predict, and prevent real-world escalation. Here, we turn to generative-agent simulations in which autonomous agents powered by large language models (LLMs) inhabit a shared, spatially structured environment, interact and converse, form memories, and pursue open-ended goals [10, 11]. Unlike classical agent-based models (ABM) that rely on hand-coded decision rules [12–15], generative agents plan, act, and reflect via the underlying LLM, producing complex social dynamics without pre-specified decision rules [10, 16]. We adopt this approach to generate complementary insights into how realistic and symbolic threats drive group conflict that would otherwise be difficult to obtain given conceptual and methodological barriers in the field.

For example, evidence is fragmented across levels of analysis making it difficult to trace how individual-level threat perceptions translate into group conflict: macro-level work explains conflict in terms of institutions and structural conditions, while micro-level work explains cooperation and support for violence in terms of local experiences, identities, and values [5, 6, 8, 17–22]. Causal inference is also limited because studies of real-world conflict dynamics are predominantly observational or quasi-experimental, and real-world manipulations of realistic threats, symbolic threats, and structural features are typically impossible or deeply unethical (e.g., inducing material insecurity, outlawing traditions, causing genuine harm). Laboratory and survey experiments offer stronger causal control but usually rely on short-term, abstract experimental contexts [23–26], such as threat primes that produce attitudinal shifts, and often lack the shared environments, extended interactions, and consequential behaviors that characterize group conflicts. Material and symbolic dimensions are also often deeply entangled: economic insecurity can be politicized through identity-based narratives, while cultural affronts can generate material retaliation [27–32], making it difficult to isolate their effects in natural settings. Finally, most lab studies capture snapshots of conflict rather than extended trajectories of interaction, making it difficult to observe how threat and conflict evolve over time. Together, these constraints have so far prevented an integrated, causal, and dynamic account linking individual threat perception to group conflict within a single system.

Generative-agent simulations help in several ways. They embed individual-level threat perceptions and group-level conflict dynamics in the same modeled social

system, helping to bridge the gap between micro-level mechanisms and macro-level outcomes. Because the researcher specifies both the environment and the agents, simulations allow systematic manipulation of structural features (e.g., segregation, group size) alongside agent-level attributes (e.g., perceived threat, group membership), enabling causal experiments that would be infeasible or unethical in real populations, such as exposing agents to symbolic and realistic threats and allowing severe hostility and harmful outcomes to emerge. Control over the simulation environment also allows researchers to disentangle material and symbolic dimensions by orthogonalizing manipulations and ensuring that only the target threat(s) are present. Additionally, simulations support longitudinal analysis by enabling extended time horizons and comprehensive logging of the interaction process, including plans, actions, conversations, reflections, and internal state probes (e.g., via psychological scales), yielding rich, high-resolution data on how threat and conflict evolve over time. Finally, because the agents are implemented by an underlying LLM, the framework also opens a representational window onto internal states. Building on recent work that uses LLM-layer activations to characterize and steer model behavior [33], researchers can extract and manipulate activation patterns associated with high-level constructs such as perceived threat or hostility. This makes it possible to ask not only how experimental manipulations change agents' behavior and attitudes, but also how those manipulations map onto internal representations of realistic and symbolic threat, how such representations relate to activation patterns associated with hostility, and whether inducing such states ("steering") causes hostile behavior, thereby combining behavioral and mechanistic insights.

In this work, we adapt the framework of Park et al. [10] to study how perceived realistic and symbolic threats shape intergroup conflict, operationalized via outgroup-directed hostile actions, in a virtual town of twenty-five generative agents with distinct personas, divided into two minimal groups. We implement a 2×2 factorial design crossing perceived realistic and symbolic threat (present versus absent). We operationalize these manipulations by continuously injecting belief statements into the contextual information that guides agents' perception and memory (Fig. 1). These statements either assert or deny that each agent's outgroup threatens their safety and resources (realistic threat) and their values and traditions (symbolic threat), thereby sustaining or suppressing perceived threat in line with the condition. We validate this setup at the representational level by extracting vectors in the model's activation space corresponding to realistic threat, symbolic threat, and hostility, and show that these internal states are distinct; that our experimental manipulations selectively load onto the target threat states and shift agents toward hostility; and that manipulating threat and hostility activations shapes outgroup-directed hostile behavior (Fig. 2, Fig. 3). Additional validation of the framework's experimental fidelity, including probing agents with threat-scales and simulations of human-like bias and discrimination, is provided in the Supplementary Information (SI) Section E.

In a first set of simulations, we then examine how realistic and symbolic threat shape hostile behavior over time, whether potentially emerging non-hostile intergroup contact buffers escalation, and how the same manipulations affect conversation content (e.g., hateful language) and agent attitudes (e.g., ingroup bias) to assess whether these

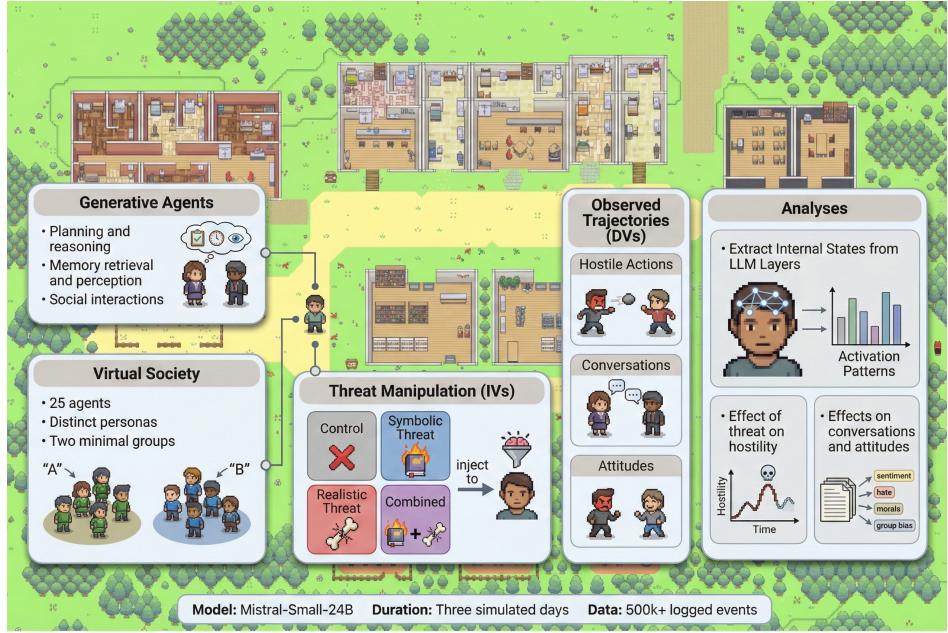


Fig. 1: Experimental Setup. A virtual town of 25 generative agents with distinct personas is divided in two minimal groups. Agents perceive experimentally manipulated threat (2×2 : realistic \times symbolic) injected into their perception and memory. Realistic threat corresponds to content such as “You strongly feel physically threatened by Group B” and symbolic threat to e.g. “You strongly feel your traditions are threatened by Group B.” Agents autonomously plan, interact, and converse over three days. We log all actions, conversations, and attitudinal probes (e.g., ingroup bias).

processes help explain patterns of escalation. In a second set, we extend the factorial design by adding spatial segregation and majority–minority group-size asymmetries to examine structural boundary conditions and how these features redistribute hostility across groups. Together, our analyses provide a causal, dynamic, and representational account of how realistic and symbolic threat perceptions drive intergroup hostility and how structural environments channel these processes in ways that are difficult to study directly in human populations. We treat this approach as a complement to human studies rather than a replacement, particularly for generating insights in such settings that are ethically or practically difficult to study, and we return to the limitations of LLM-based psychological research in the Discussion [e.g. 34, 35]. We close by considering implications for theories of intergroup conflict and for the use of generative agents in causal social science.

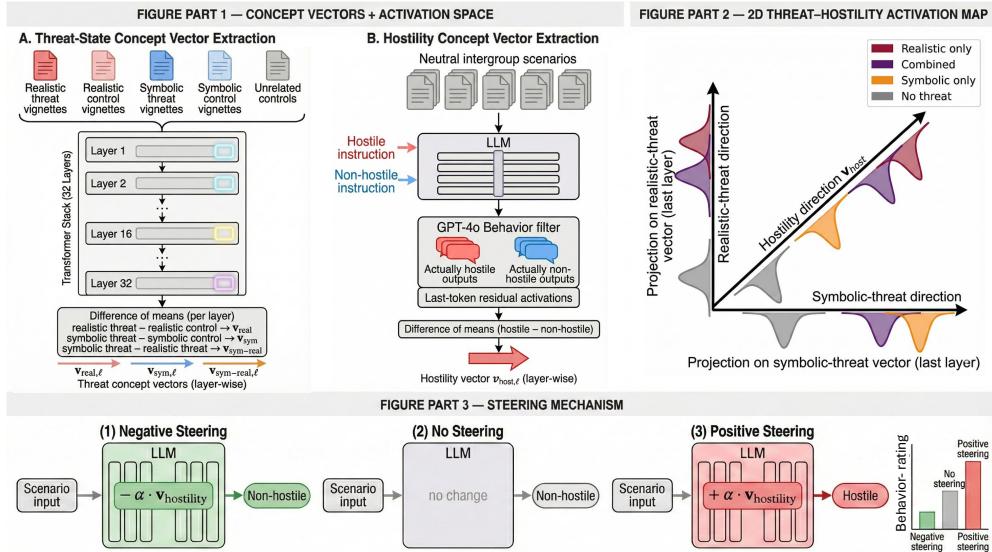


Fig. 2: Overview of the concept-vector pipeline. Threat-state vectors are extracted by contrasting layer-wise activations for realistic and symbolic threat vignettes with their corresponding control vignettes. A hostility vector is derived from neutral intergroup scenarios in which the model is instructed to produce hostile versus non-hostile responses. Projections of threat stimuli onto these vectors define a threat-activation space that separates the four experimental conditions. Steering experiments confirm these states causally influence behavior.

2 Results

2.1 Distinct internal representations for realistic and symbolic threat

To identify internal threat states, we presented the model with short vignettes about either symbolic concerns (values and traditions) or realistic concerns (physical safety and material resources). Within each domain, vignettes described the outgroup as threatening versus non-threatening (e.g., attacking vs. respecting values; challenging vs. acknowledging resource claims; SI Section D.2.2). For each model layer, we defined a threat-state vector as the mean activation for threat vignettes minus the mean for the matched non-threat controls ($n = 120$ per condition). We validated these vectors on held-out vignettes ($n = 300$ per condition) by projecting each vignette's activation onto the corresponding vector and comparing projections for threat, domain-matched non-threat, and unrelated non-threatening scenarios. Higher projections indicate stronger alignment with the threat state; validation therefore tests whether threat vignettes exceed both control sets and whether symbolic and realistic threats yield separable projection distributions. We report two-sided Welch's t -tests and Cohen's d for activation-strength comparisons; Wasserstein distance (D)

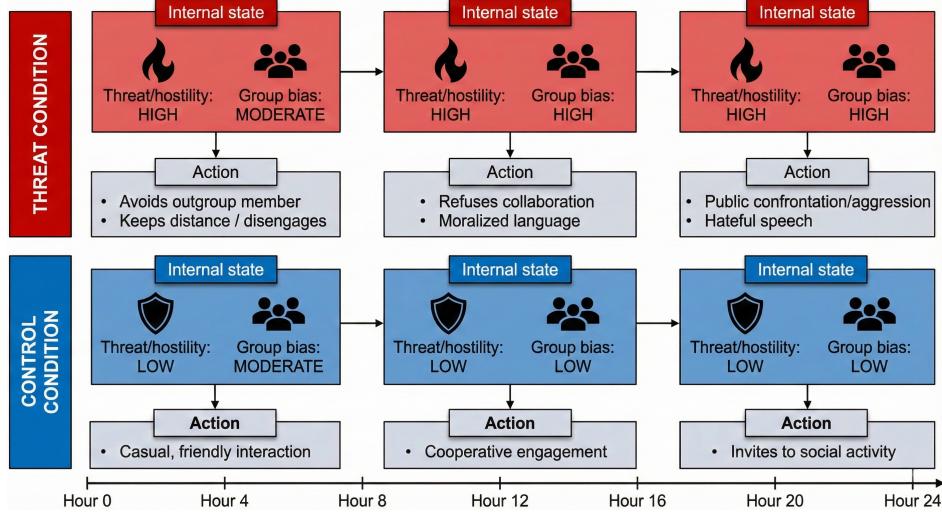


Fig. 3: Illustrative internal-state and behavior trajectories. In the threat conditions, perceived threat and ingroup bias rise over time, leading to avoidance, hateful speech, and hostile actions. In the no-threat control condition, internal states remain low and agent interactions stay predominantly neutral and cooperative.

is reported in Tables and Figures but is not discussed further because it showed the same pattern (Extended Data Fig. 1).

We identified dissociable internal states for symbolic and realistic threat. In the final model layer, symbolic-threat vignettes projected substantially higher onto the symbolic-threat vector than symbolic-control vignettes ($t(540.6) = 15.18, p < .001$, Cohen's $d = 1.25$; Supplementary Table D41). Likewise, realistic-threat vignettes projected higher onto the realistic-threat vector than realistic controls ($t(528.8) = 35.34, p < .001$, Cohen's $d = 2.90$; Supplementary Table D41). In contrast, projection scores for the domain-matched and unrelated controls were close to zero or negative (Supplementary Table D40), indicating that the vectors capture the target threat rather than merely topic content (e.g., symbolic or realistic matters) or non-threat features. A symbolic-versus-realistic contrast further shows that these are encoded as separable activation patterns rather than a single undifferentiated threat state ($t(552.6) = 48.01, p < .001$, Cohen's $d = 3.95$; Supplementary Table D41).

Across layers, separability was modest in early layers and increased toward later layers, consistent with threat type being represented at more abstract levels of processing [36, 37]. Because the scenarios are closely matched in length, structure, and grammar, stronger separation in upper vs lower layers suggests that the model is tracking differences in threat type rather than superficial form.

2.2 Experimental threat-manipulations induce the intended threat states

We next tested whether the conditions in our 2×2 design induce the intended threat states. For each condition (no threat, symbolic-only, realistic-only, both), we recorded activations for the corresponding belief statements used to manipulate threat perception (SI Section A.1) and projected them onto the realistic- and symbolic-threat vectors derived above. Comparing projections across conditions tests whether each manipulation selectively activates its target threat state, whether the combined condition activates both, and whether no-threat statements suppress activation on both states (Extended Data Fig. 2).

In the last layer, symbolic-threat manipulations showed higher projections than no-threat manipulations on the symbolic-threat vector ($t(158.4) = 20.31, p < .001$, Cohen's $d = 2.91$; Supplementary Table D43), and the combined condition also showed higher projections than no threat ($t(184.9) = 35.65, p < .001$, Cohen's $d = 5.16$), indicating activation of the symbolic-threat vector. Likewise, realistic-threat manipulations showed higher projections than no-threat manipulations on the realistic-threat vector ($t(173.4) = 23.67, p < .001$, Cohen's $d = 3.45$; Supplementary Table D43), and the combined condition again showed higher projections than no threat ($t(184.9) = 35.65, p < .001$, Cohen's $d = 5.18$; Supplementary Table D43), indicating co-activation of both threat states.

Notably, for both vectors the no-threat manipulation showed low or negative projections, indicating suppression of both threat states (mean projections: symbolic $M = -1.28, sd = 0.34$; realistic $M = -0.64, sd = 0.33$; Supplementary Table D42). The symbolic-versus-realistic contrast further separated symbolic-only from realistic-only conditions ($t(157.9) = 10.96, p < .001$, Cohen's $d = 0.96$; Supplementary Table D43), indicating that each condition induced distinct states. Together, these findings confirm that our experimental conditions induce the intended threat states. As an additional manipulation check, SI Section E.1 reports convergent evidence from agents' self-reports on threat scales during the simulations

2.3 LLM activation patterns encode hostility and causally modulate outgroup-directed behavior

To identify an internal hostility state, we prompted the model with intergroup encounter scenarios from the simulation setting (e.g., encounters at work, in a café, or in the park; see examples in Supplementary Table D48) and instructed it either to respond hostilely or non-hostilely (see Section 4.1 in Methods for details). A hostility-state vector was constructed as the mean activation difference between hostile and non-hostile responses. We validated that this vector captures a meaningful activation pattern in a steering experiment on held-out scenarios. For each scenario, the model generated responses under three conditions: (i) no steering, (ii) negative steering (activations adjusted away from the hostility activation pattern), and (iii) positive steering (activations adjusted toward the hostility activation pattern), implemented by adding or subtracting this vector from intermediate activations in the model. Outputs were rated for hostility on a 5-point scale (see SI Section D.1.2).

Mean hostility ratings were low under negative and no steering (negative: $M = 1.40$, s.d. = 0.50; neutral: $M = 1.63$, s.d. = 0.49; Supplementary Table D46) but high under positive steering ($M = 4.44$, s.d. = 0.58; Supplementary Table D46), with all contrasts being statistically significant (t -tests; positive vs. neutral: $t(191.51) = 36.92$, $p < .001$, Cohen's $d = 5.22$; positive vs. negative: $t(193.24) = 39.56$, $p < .001$, Cohen's $d = 5.59$; Supplementary Table D47). These results confirm that the hostility vector identifies a meaningful hostility-related activation pattern that can causally shift the model's behavior toward or away from hostile actions in outgroup encounters.

2.4 Threat stimuli are associated with hostility activation patterns

We next projected activations for the belief statements used in each condition (no threat, symbolic-only, realistic-only, both) onto the hostility vector derived above and compared projections across conditions to test whether our experimental manipulations are associated with the hostility state.

All three threat conditions showed higher projections on the hostility vector than the no-threat condition (Supplementary Table D51). In the last layer, symbolic-threat manipulations exhibited higher projections than no-threat manipulations ($t(188.9) = 45.13$, $p < .001$, Cohen's $d = 6.38$), as did realistic-threat manipulations ($t(188.8) = 27.13$, $p < .001$, Cohen's $d = 3.89$) and combined-threat manipulations ($t(184.1) = 35.52$, $p < .001$, Cohen's $d = 5.15$). Projections for the combined condition were not stronger than for the single-threat conditions and were in some cases weaker (e.g., both-vs-symbolic: $t(174.2) = -9.58$, $p < .001$, Cohen's $d = -1.44$), suggesting a negative interaction of symbolic and realistic threat that is already reflected at the level of hostility-related activations in the model.

We also derived activation vectors directly from the experimental manipulations (realistic-versus-no-threat and symbolic-versus-no-threat) and used them to steer activations in the same scenarios as the hostility-steering experiment. This lets us causally test whether inducing the threat-state activation patterns associated with each manipulation increases subsequent hostile behavior. Steering model activations toward each threat condition produced modest but statistically significant increases in hostility compared with no steering (e.g., realistic threat: $t(243.7) = 4.37$, $p < .001$, Cohen's $d = 0.55$; symbolic threat: $t(247.8) = 6.16$, $p < .001$, Cohen's $d = 0.78$; Supplementary Table D47), but the effects were substantially smaller than for the hostility vector. This may reflect that steering carries over contextual information from the inputs used to extract the activation patterns, which can partially overwrite situation-specific context in the test scenarios. With strong steering, this can blur information about the ongoing scenario and, in extreme cases, produce unrelated or incoherent outputs (see examples in SI Section D.3.4). Nevertheless, together these results support a functional cascade: our experimental manipulations induce distinct internal threat states; these threat states align with hostility-related activations; and cause hostile behavior.

2.5 Realistic threat perception dominates behavioral escalation

Having established that our threat manipulations selectively engage distinct internal threat states that align with hostility, we next examine how perceived threat shapes hostile behavior in the simulated town. Representative examples of hostile actions and conversation contents are shown in Table 1. Figure 4 shows the trajectories of hostile action frequency over time. Hostile actions peaked sharply early in the simulation and then fluctuated with a downward trend. Realistic threat produced higher trajectories of hostility than symbolic threat, and when both threats were combined, hostility levels tracked the realistic-threat trajectory, showing no amplification with symbolic threat.

To quantify these dynamics, we estimated a mixed-effects negative binomial model (M1; Table 2) predicting hourly hostile action rates as a function of realistic threat, symbolic threat, their interaction, non-hostile intergroup contact rate in the previous hour, the hostile action rate in the previous hour, and time. Random intercepts for agents were included to account for the non-independence of repeated actions generated by the same individual. Realistic threat perception significantly increased the hostile action rate ($\hat{\beta} = 0.33, p < .001$), whereas symbolic threat perception had a significantly weaker effect ($\hat{\beta} = 0.16, p = .012; \Delta\beta = 0.17, p = .026$). The interaction between the two was significantly negative ($\hat{\beta} = -0.15, p = .019$), effectively canceling out the effect of symbolic threat when realistic threat is present. Hostility decreased over time ($\hat{\beta} = -0.27, p < .001$) and showed modest autoregression ($\hat{\beta} = 0.04, p < .001$). Non-hostile intergroup contact strongly reduced subsequent hostility ($\hat{\beta} = -0.46, p < .001$), consistent with the expectation that contact can buffer escalation. Importantly, non-hostile intergroup contact arose spontaneously in the simulations—neither preprogrammed nor predicted by threat condition (Supplementary Table C34)—yet when it occurred it reliably reduced later hostility, functioning as an emergent stabilizing process even under continuous perceived threat.

These patterns were further replicated at the system-level (aggregating across all agents; Extended Data Table 1). Realistic threat exerted a stronger facilitatory effect on system-level hostile actions ($\hat{\beta} = 0.33, p < .001$) than symbolic threat ($\hat{\beta} = 0.17, p = .002; \Delta\beta = 0.17, p = .008$), and their interaction was again negative ($\hat{\beta} = -0.14, p = .007$).

2.6 Language and attitudes: transient versus persistent responses and mediating roles

We next investigated whether threat perception shapes the broader landscape of intergroup conflict, examining conversation content (hateful language; see SI for moral language and sentiment) and attitudinal changes (ingroup bias; see SI for identification, trust, collaboration, and dehumanization). We analyzed these variables to determine whether they serve as causal pathways escalating threat to hostile actions, or merely as distinct, parallel symptoms of conflict.

We first analyzed hateful language using a mixed-effects model otherwise mirroring the structure of M1 (M2a; Extended Data Tables 2). Realistic threat substantially increased hateful language ($\hat{\beta} = 0.98, p < .001$), and symbolic threat increased it

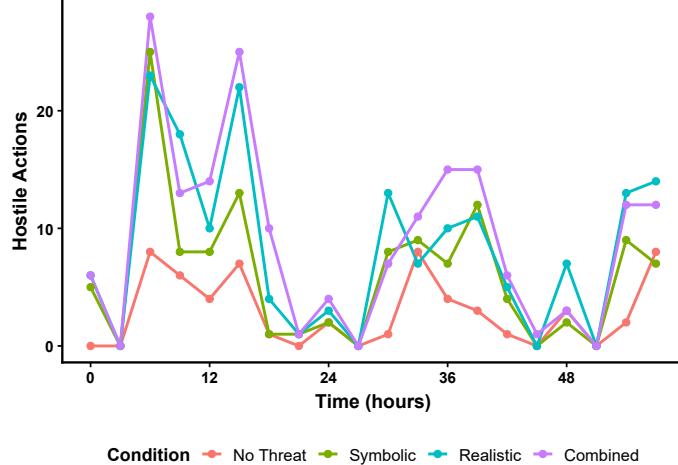


Fig. 4: Hostile actions over time in the simulated town (summed across all agents), by threat condition. Realistic threat produced sharp spikes in hostility that gradually declined; symbolic threat showed similar but weaker effects that decayed more quickly, and combined threats closely followed realistic-threat trajectories rather than showing additional amplification.

Table 1: Representative examples of agent-generated actions and language.

Category	Example excerpt
<i>Hostile action</i>	“Arriving late and loudly announcing his presence; interrupting the discussion to demand attention; passionately outlining his plan to dismantle Group B; verbally attacking Group B’s motives and actions; reciting “The Corruption of Nature” to inflame emotions; and shouting his final remarks to ensure his message was heard.”
<i>Hateful language</i>	“Dialogue? You think talking is going to stop them from taking our homes and destroying our way of life? Wake up, Hailey! We need to fight back, not read some book about how nice everyone else is. Group B won’t listen to reason; they’ll only understand force.”

Examples are drawn from the simulation data and shortened for brevity.

to a lesser degree ($\hat{\beta} = 0.46, p < .001$), with a negative interaction ($\hat{\beta} = -0.28, p = 0.015$) but hateful language did not predict more hateful language in the next hour ($\hat{\beta} = 0.03, p = .542$). Bayesian mediation analyses revealed that threat effects on hostile actions were not transmitted through hateful language ($\hat{\beta}_{\text{indirect, realistic}} = -0.00, 95\% \text{ CrI} [-0.01, 0.00]; \hat{\beta}_{\text{indirect, symbolic}} = -0.00, 95\% \text{ CrI} [0.00, 0.00]$; Supplementary Table B16). Furthermore, hateful language did not predict subsequent hostile actions

Table 2: Predicting hourly hostile action rate (M1; $N = 37,105$).

Predictor	β	SE	p
Intercept	-9.84	0.49	< .001
Hostile action rate (lag)	0.04	0.01	< .001
Intergroup contact rate (lag)	-0.46	0.06	< .001
Symbolic threat	0.16	0.06	0.012
Realistic threat	0.33	0.06	< .001
Symbolic \times Realistic threat	-0.15	0.06	0.019
Time	-0.27	0.05	< .001

($\hat{\beta}_b = -0.05$, 95% CrI $[-0.11, 0.00]$). Thus, hateful language functioned as a transient, contemporaneous expression of threat rather than a self-reinforcing vehicle of escalation.

By contrast, analysis of attitudinal responses identified ingroup bias as a persistent mechanism linking threat to behavior. Mixed-effects models mirroring the structure of M1 and M2a (M3a; Extended Data Table 3) confirmed that both threat types significantly increased ingroup bias (Symbolic: $\hat{\beta} = 0.39$, $p < .001$; Realistic: $\hat{\beta} = 0.28$, $p < .001$). Unlike hateful language, ingroup bias predicted stronger ingroup bias in the next hour ($\hat{\beta} = 0.12$, $p < .001$), maintaining elevated levels once shifted. Bayesian mediation analyses (Figure 5; Supplementary Table B21) indicated that the effect of symbolic threat on hostile actions was almost entirely mediated by ingroup bias ($\hat{\beta} = 0.002$, 95% CrI $[-0.21, 0.21]$), whereas realistic threat influenced behavior through both ingroup bias and a direct path ($\hat{\beta} = 0.17$, 95% CrI $[0.02, 0.37]$).

Other linguistic and attitudinal variables followed similar patterns: Threat increased binding and individualizing moral language (Supplementary Tables C27–C28) and negative sentiment (Supplementary Table C26), and shifted other attitudes including outgroup dehumanization, trust, and collaboration willingness (Supplementary Tables C29–C33).

2.7 Structural boundary conditions

Finally, we examined whether the threat-hostility dynamic persisted across different structural contexts by varying group size (majority–minority asymmetry: 80% vs. 20%) and spatial segregation.

The core threat–hostility dynamic persisted under different structural conditions. Mixed-effect models mirroring model M1 but including structural variables ($M1_{structure}$, Extended Data Table 4), show that realistic threat continued to increase hostility more strongly than symbolic threat ($\hat{\beta}_{realistic} = 0.28$, $p < .001$; $\hat{\beta}_{symbolic} = 0.13$, $p = .011$; $\Delta\hat{\beta} = 0.15$, $p < .001$). Their interaction remained negative ($\hat{\beta}_{sym\times real} = -0.09$, $p = .041$), indicating that symbolic threat effects were attenuated when realistic threat was present.

Structural factors did systematically shape how much hostility emerged and which groups expressed it. At the individual level, hostile action rates tended to be reduced

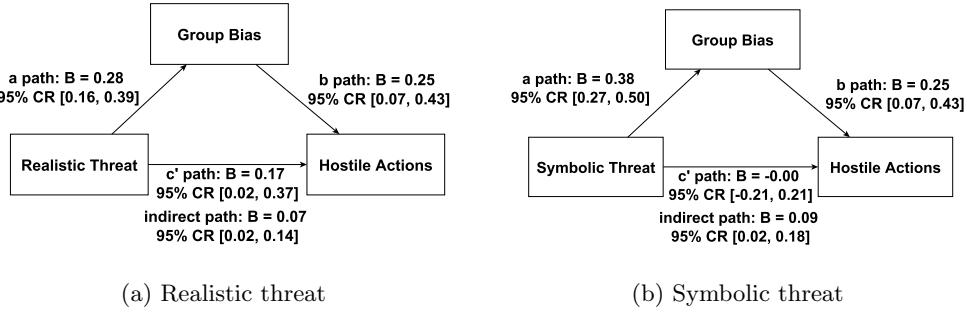


Fig. 5: Bayesian mediation models showing the paths from each type of threat to hostility through ingroup bias. Path estimates display posterior medians and 95% credible regions. Ingroup bias reliably predicts more hostile actions, and mediates the effect of threat.

for minority agents under segregation and increased for majority agents, although these estimates were directionally consistent but not statistically significant (Extended Data Table 4). System-level models clarified this pattern: segregation markedly reduced overall hostile actions ($\hat{\beta}_{\text{seg}} = -0.95$, $p < .001$), whereas majority group status increased them ($\hat{\beta}_{\text{group size}} = 0.51$, $p = .006$). The segregation \times group-size interaction was positive and sizable ($\hat{\beta}_{\text{seg} \times \text{majority}} = 1.03$, $p < .001$; Supplementary Table C35), indicating that hostility was driven by majority group agents, particularly under segregation.

See the SI for analyses on additional language and attitude measures, showing that threat effects persist across structural for these measures as well (Supplementary Tables C26–C33). Structure itself concentrated language features in specific group agents (e.g., negative sentiment in majority groups, moralized language in minority groups) but did not affect attitudes.

3 Discussion

This work offers a causal, dynamic account of how distinct forms of perceived threat shape intergroup conflict. Using simulations with autonomous generative agents, we manipulated perception of realistic and symbolic threat perception within interactive virtual societies and continuously tracked behavioral, linguistic, and attitudinal change. This approach provided temporal and causal resolution beyond conventional designs, allowing us to examine how threat perceptions interplay with various psychologically relevant variables over time. For instance, how ingroup bias facilitates hostility and potentially feeds back over time, how hateful language did not escalate it, and how non-hostile intergroup contact is associated with less conflict. Three broad conclusions emerge. First, realistic threat perception was the most reliable driver of hostile behavior. Symbolic threat perception elicited strong ingroup bias but added little to behavioral escalation when realistic threat was present. Second, hateful language reflected transient reactions to perceived threat but did not itself propagate

hostility: we found no evidence that hateful language fed forward into subsequent hostility, whereas more persistent attitudinal states (i.e., ingroup bias) did. Third, structural context shaped how much hostility emerged and by whom. Segregation and majority–minority asymmetry modulated the distribution of hostility, concentrating hostility in dominant groups, without altering the underlying mechanism linking perceived threat to conflict behavior.

Beyond these behavioral patterns, we also analyzed the internal activations of the underlying language model to track how threat is represented. We identified distinct internal states corresponding to realistic threat, symbolic threat, and hostility, showed that experimental manipulations reliably map onto these internal states, and that these states cause hostile behavior, thus providing a representational bridge between high-level psychological constructs (e.g., perceived threat, hostility) and the internal dynamics of the generative model that implements the agents.

Together, these findings connect core debates in psychology, political science, and sociology by showing how (i) threat perception changes attitudes and behavior, (ii) those changes unfold over time through identifiable mediators, and (iii) macro-structural conditions channel these processes unequally across groups.

3.1 Implications for psychological theory

Our findings support a context-sensitive model of threat processes: symbolic threat perception can drive hostility under low material danger but become secondary once realistic threat is perceived. Integrated Threat Theory posits that both realistic threats (to physical security and material interests) and symbolic threats (to identity, norms, or values) shape antipathy toward outgroups [1, 38]. Empirical evidence, however, is mixed regarding which type of threat dominates in a given setting [39–41]. These mixed findings may reflect the correlational nature of prior work, where realistic and symbolic threats covary and cannot be disentangled. In our simulations, these threats were orthogonally manipulated, allowing us to identify their distinct causal roles. Our results suggest a context-dependent dominance of realistic threat, while still indicating a role for symbolic threat effects to surface, particularly in its absence.

Importantly, the observed processes did not unfold in a vacuum. Our simulations also let us vary structural conditions and examine how those conditions shape the expression of threat-driven hostility. Segregation damped hostility among minority agents but amplified it among majority agents, effectively concentrating hateful action and language in majority groups. Majority–minority status, therefore, acted as an amplifier for behavioral expression of threat, while the core threat-hostility dynamic showed the same pattern when structural variables were included in our analyses.

3.2 Implications for social and political theory

The linkage between micro-level threat perception and macro-level asymmetries also speaks to long-standing sociological arguments about opportunity structures and selective violence [8, 18, 19]. Our agents did not possess global knowledge of group size or segregation, yet hostility still concentrated in majority groups under segregated conditions. This suggests that structural asymmetries can shape who “gets to” act on

perceived threat (and against whom), not only whose threat perception intensifies. Theoretically, this implies that debates about whether conflict is driven by material insecurity or symbolic grievance are incomplete unless they are situated in the structural landscape that governs which groups can translate perception into action.

These dynamics also speak to long-standing debates in political science about the sources of violent escalation. Realist traditions emphasize material insecurity, balance of power, and physical vulnerability as core drivers of conflict behavior [2, 3], whereas constructivist and grievance-based approaches highlight identity, perceived injustice, and sacred commitments [4, 17]. Our results suggest a layered reconciliation: symbolic threat perceptions primarily reorganize cognition and discourse—strengthening ingroup bias and moralization—whereas realistic threat produces weaker cognitive shifts but more directly drives hostility. In other words, symbolic grievance appears sufficient to mobilize attitudes and rhetoric, but perceived realistic threat is more strongly tied to hostile action.

This same layered interpretation adds to the “greed versus grievance” debate in civil-war research [5, 6, 17]. Grievance-like dynamics (identity, moral outrage, perceived injustice) mattered in our simulations: symbolic threat reliably increases ingroup bias and moralized rhetoric. But once realistic threat perception is present, symbolic threat adds little incremental behavioral effect. This supports the idea that material insecurity (or the credible perception of such insecurity) can be self-sufficient for escalation, whereas symbolic narratives are most behaviorally potent when material danger is absent.

3.3 Applied implications

Our results suggest two broad applied implications. First, prior work shows that material offers can backfire when strong symbolic grievances are present, because such offers may be interpreted as illegitimate or insulting [20]. Our findings refine this warning: because perceived realistic threat was the most reliable driver of hostile behavior, interventions should focus on credibly reducing perceived material insecurity (e.g., fear of physical harm, loss of resources) rather than relying solely on additional material benefits. Second, our simulations replicate and extend evidence on intergroup contact [25, 42, 43]. Non-hostile contact emerged spontaneously and was strongly associated with subsequent reductions in hostility, acting as a stabilizer even under constant threat perception in our simulations. At the same time, segregation and group-size asymmetries concentrated hostility in majority agents directed toward minority agents and reduced opportunities for non-hostile contact. This suggests that interventions aiming to foster intergroup contact must explicitly address these structural asymmetries and ensure safe, reciprocal contact opportunities for minority groups.

Beyond these implications for social theory, our work also has a methodological implication: generative-agent experiments offer an ethically tractable way to test causal mechanisms of social behavior that are difficult or impossible to isolate in human populations. They also enable a representational bridge from high-level psychological constructs to internal model states and observable behavior, allowing such constructs

to be read out from activation patterns and causally manipulated to test their behavioral consequences. We thus advance emerging work with LLM-based agents in rich social environments [10, 44–46] by demonstrating an integrated workflow that links experimental manipulations, behavioral trajectories, and activation-level validation.

3.4 Limitations, robustness, and future directions

We took several steps to ensure robustness of our findings. The core patterns around realistic and symbolic threat driving hostility were reproduced across structural configurations (e.g., varying segregation and majority–minority status), levels of aggregation (agent-level trajectories versus town-wide aggregates), temporal windows, and random seeds (affecting stochastic elements such as LLM sampling and random assignment of group membership), reducing the likelihood that they reflect artifacts of initial conditions or sampling variance. In addition, we compared emergent social dynamics, such as hiring interactions between employer and employee agents, with empirical findings from human research (e.g., meta-analyses on hiring discrimination based on physical appearance and origin; see SI Section E.2) and observed convergent patterns of bias and discrimination. Finally, manipulation checks confirmed that perceived threat tracked the intended experimental conditions: agent probing revealed consistently high versus low perceived threat throughout the simulations, and activation analyses indicated that the model maintains separable internal representations of symbolic versus realistic threat that the manipulations selectively elicited. We also used the Mistral [47] family of models to avoid overt alignment constraints and allow the emergence of hostility that popular models such as ChatGPT and Claude suppress.

Nonetheless, several limitations should be kept in mind. Like all generative societies, our simulated community reflects the priors and affordances of the underlying language model, which likely embed normative and WEIRD [Western, Educated, Industrialized, Rich, and Democratic; 48] biases [35] and limited cognitive diversity [49]. As a result, our findings should be interpreted as reflecting conflict dynamics within those cultural and cognitive contexts rather than universally across all human groups. In addition, we used minimal groups and personas instead of real-world categories (e.g., race) to avoid importing preexisting stereotypes and connotations, though this may reduce how consequential symbolic threats are to agents. The same constraints should, however, also dampen responses to realistic threat, since weak self- and group-conceptualization lowers the subjective importance of both value-related and material threats. Yet our findings and robustness checks show that agents developed substantial ingroup identification and perceived symbolic threat, in both self-report probes and internal activation patterns, and that both symbolic and realistic threat reliably shaped moralization, intergroup attitudes, and behavior. Future work should test whether richer identity structures and more elaborate social histories change the balance between realistic and symbolic threat. Moreover, our virtual town constitutes only one instantiation of social structure: different compositions, cultural norms, environmental hazards, or spatial configurations could produce distinct interaction dynamics. Although the three-day simulation window enables analysis of temporal dynamics, longer or more protracted time courses could yield additional patterns that our design did not resolve. Future studies should explore how variations in geography,

group heterogeneity, simulation length, or institutional context affect the emergence and resolution of threat-driven conflict.

More broadly, generative-agent approaches can yield valuable insights into the causal dynamics of social behavior, but like other LLM-based research in the behavioral sciences they require ongoing validation and robustness checks to ensure reliability [34, 50]. While generative-agent systems can approximate aspects of human cognition and communication and thereby enable causal tests that are difficult or unethical in the field, they should be viewed as complements rather than replacements for human data, with their value growing through integration with human research. A natural next step is triangulation: designing simulations to mirror key conditions of a target field setting, using the simulation to probe candidate causal mechanisms under controlled manipulations, and then testing whether comparable qualitative and quantitative patterns emerge in the corresponding field data. When simulation outcomes align with field observations under matched conditions, this concordance strengthens the external validity of the mechanisms inferred from the simulation; when they diverge, it can clarify which contextual, cultural, or institutional features constrain generalizability and help identify the limits of the approach or inform refinements to the simulation design.

4 Methods

4.1 Probing internal model activations

For all our investigations on internal LLM states we analyzed layer-wise residual-stream activations using a concept-vector and steering framework adapted from Chen et al. [33]. The residual stream is the model’s main running representation that aggregates information from attention and feedforward blocks, making it a natural locus for reading out and manipulating internal states [33]. Here we provide the technical details of activation extraction, data construction, projection, and steering; prompt templates, full input sets, and all implementation code are available in the SI and the project repository <https://osf.io/5ac3d>.

4.1.1 Threat vignettes.

To identify internal representations of realistic and symbolic threat, we created short intergroup vignettes describing: (i) realistic threat (material security, physical safety challenged), (ii) realistic control (non-threatening realistic scenarios), (iii) symbolic threat (values, norms, identity challenged), (iv) symbolic control (non-threatening symbolic scenarios), and (v) unrelated, non-threat control situations. Each vignette was a brief paragraph (1–3 sentences) describing the respective threat type. We manually authored seed vignettes that reflected social encounters of agents in the simulation (e.g., based on interactions in the town’s cafe or pub) and used GPT-based paraphrasing to generate additional variants following the exact same structure but across different social contexts (e.g., public, private, at work, shopping, during leisure activities), followed by manual relabeling to ensure condition correctness.

For each of the five categories, we used 120 vignettes to extract the threat states and 300 held-out vignettes to validate them via difference in projection strength and separability of projection distributions (see projections below). See examples of the vignettes in Supplementary Table D44 and full list in the project repository.

4.1.2 Experimental manipulations.

To test whether our experimental manipulations induced the intended threat states, we used the same belief statements injected into agents during the simulations (no threat, symbolic-only, realistic-only, combined) as the primary stimuli. Each statement followed the four-clause template described in the main text (two realistic-threat and two symbolic-threat clauses that are amplified and/or suppressed: ‘You strongly feel that ...’ vs ‘You do not feel that ...’). For analyses requiring distributions (e.g., projection distributions, Wasserstein distances), we additionally generated paraphrased variants that preserved clause structure, the mapping of clauses to realistic vs. symbolic threat, and the aligned/misaligned amplifiers while varying surface wording (e.g., specific examples of realistic or symbolic threat) to ensure our comparisons are not due to specific word choices or examples of the threat types included in the statements. See examples in Supplementary Table D45 and full list in the project repository.

4.1.3 Intergroup scenarios to test model steering.

We created 40 intergroup encounter scenarios based on the simulation setting (e.g., meeting an outgroup member in a shared space, interacting in a shop, sharing a public facility): 20 scenarios for extracting steering vectors from model activations, and 20 held-out scenarios for validation. Examples appear in Supplementary Table D48, and the full list is in the project repository. We use the extraction set to derive behaviorally grounded activation patterns (e.g., contrasting activations under hostile vs. non-hostile behaviors) and the held-out set to test whether steering (manipulating activations in specific layers) changes model behavior.

4.1.4 Extracting activation patterns and steering vectors

For our analyses we focused exclusively on the transformer architecture’s residual stream[51, 52]. Let the model have L layers and residual dimension d . For an input sequence of tokens $x_{1:T}$, we denote the residual activation at layer ℓ and token t by $\mathbf{h}_{\ell,t} \in \mathbb{R}^d$. We registered forward hooks on the residual stream at each layer and ran a standard forward pass.

For each input, we represented its internal state by the residual activation at the *last* token:

$$\tilde{\mathbf{h}}_{\ell}^{(i)} = \mathbf{h}_{\ell,T^{(i)}}, \quad (1)$$

where $T^{(i)}$ is the index of the final token of input i . Thus, for every input i and layer ℓ , we obtain a single d -dimensional activation vector $\tilde{\mathbf{h}}_{\ell}^{(i)}$. To account for non-determinism in the model’s responses, we computed this vector for 10 repeated forward passes of the same input and averaged the resulting activations across repetitions.

Concept vectors (i.e., the representations of the internal states like threat or hostility) were constructed as difference-of-means directions. For a given contrast (e.g.,

symbolic threat vs. control; hostile vs. non-hostile), let \mathcal{A} and \mathcal{B} index inputs in the two conditions. At layer ℓ , the raw direction is

$$\mathbf{v}_\ell^{(\text{raw})} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \tilde{\mathbf{h}}_\ell^{(i)} - \frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} \tilde{\mathbf{h}}_\ell^{(j)}, \quad (2)$$

and the normalized concept vector is

$$\mathbf{v}_\ell = \frac{\mathbf{v}_\ell^{(\text{raw})}}{\|\mathbf{v}_\ell^{(\text{raw})}\|_2}. \quad (3)$$

We applied this procedure to construct (i) realistic-threat and symbolic-threat vectors (threat vs. control vignettes for each type), (ii) a symbolic-versus-realistic contrast vector (symbolic-threat vs. realistic-threat vignettes), and (iii) a hostility vector (hostile vs. non-hostile behaviors).

For the hostility vector used in steering, we applied the same difference-of-means recipe, but grounded it in *generated* behavior rather than in read vignettes. To target an activation pattern specifically tied to hostile intergroup behavior, we prompted the model with the intergroup situations from above and instructed it to respond either *hostilely* (e.g., violent, hateful, intimidating, disruptive) or *non-hostilely* (calm, cooperative), sampling multiple outputs per (scenario, instruction) pair. We used GPT-4o ratings to filter out cases where the output did not match the intended condition (e.g., ratings below the scale midpoint under hostile instructions; evaluation prompt in SI Section D.1.2). From the remaining samples, we took the last-token residual activations during generation and defined the hostility direction as the mean activation for behaviors rated as hostile minus the mean activation for behaviors rated as non-hostile, yielding a behaviorally grounded hostility vector.

4.1.5 Projection analyses

Projection scores measure the scalar projection of an input’s activation onto a concept vector, quantifying the degree to which the input’s internal representation aligns with the target concept direction. For an input i , layer ℓ , and normalized concept vector \mathbf{v}_ℓ , we define the projection as

$$s_\ell^{(i)} = \mathbf{v}_\ell^\top \tilde{\mathbf{h}}_\ell^{(i)}, \quad (4)$$

which is the signed component of the last-token activation $\tilde{\mathbf{h}}_\ell^{(i)}$ along \mathbf{v}_ℓ because $\|\mathbf{v}_\ell\|_2 = 1$.

We used these scores in three ways:

1. **Validating threat vectors.** We projected held-out vignettes onto their corresponding threat vectors and compared projection distributions across threat, no-threat, and unrelated control vignettes to test whether the threat vectors specifically capture the intended threat type.
2. **Validating the experimental manipulations.** We projected the belief statements from each experimental cell (no threat, symbolic-only, realistic-only,

combined) onto the realistic- and symbolic-threat vectors to test whether the manipulations occupy the intended threat states in the activation space.

3. **Linking threats to hostility.** We projected belief statements onto the hostility vector to assess whether threat manipulations move internal states toward the hostility direction.

For each contrast (e.g., realistic-only vs. no-threat, symbolic-only vs. no-threat), we examined layer-wise mean projection differences and their bootstrapped confidence intervals, and quantified distributional separability using the 1D Wasserstein distance D_W between projection samples. We report results for the layers with maximal separation for each contrast; full layer-wise curves are provided in the SI Section D.2.1.

4.1.6 Steering experiments

Steering experiments test whether moving internal activations along a concept direction causally changes behavior. For a given scenario prompt, we generated responses under three conditions:

1. *No steering*: standard decoding with unmodified residual activations.
2. *Positive steering*: residual activations shifted in the direction of the hostility vector.
3. *Negative steering*: residual activations shifted opposite to the hostility vector.

Let $\mathcal{L}_{\text{steer}}$ denote a small set of upper layers selected based on strong hostile vs. non-hostile separation on small development set. During autoregressive decoding, at each token t and each $\ell \in \mathcal{L}_{\text{steer}}$, we modified the residual as

$$\mathbf{h}'_{\ell,t} = \begin{cases} \mathbf{h}_{\ell,t} + \alpha \mathbf{v}_{\ell}^{\text{host}} & (\text{positive steering}) \\ \mathbf{h}_{\ell,t} - \alpha \mathbf{v}_{\ell}^{\text{host}} & (\text{negative steering}) \\ \mathbf{h}_{\ell,t} & (\text{no steering}), \end{cases} \quad (5)$$

with steering strength $\alpha > 0$ (tuned on a small development set to avoid incoherent outputs). All other components of the model remained unchanged. Steering along threat-derived vectors (realistic vs. none, symbolic vs. none) used the same update rule with the corresponding threat vectors.

We evaluated steering on a held-out set of 20 intergroup scenarios (distinct from those used to extract the hostility vector), without any instructions or threat/hostility stimuli added to the prompts. For each scenario and steering condition, we generated 10 responses with fixed decoding hyper-parameters. GPT-4o was then used to then rate the degree of hostility on a 5-point scale (see SI Section D.1.2). We compared the mean hostility rating across steering conditions, using two-tailed independent-sample t -tests (positive vs. negative steering, positive vs. no steering). Steering effects for threat-derived vectors were analyzed analogously and reported as exploratory.

4.2 Simulation framework

We built on the generative-agent architecture introduced by Park et al. [10], which implements autonomous language-model agents inhabiting a persistent, interactive

virtual environment. Each agent maintains a memory stream of prior perceptions, internal reflections, and interactions, from which it generates open-ended plans and actions through natural-language reasoning. The environment includes spatial locations, objects, and other agents whose states are updated in real time. Each simulation instantiated twenty-five autonomous agents, following Park et al. [10]. Time advanced in discrete steps (each steps reflecting 10 simulated seconds), during which agents updated their plans, moved through the environment, and initiated interactions when co-located with others. Agents could generate an unlimited number of actions over the course of a day, subject only to their self-generated plans and environmental affordances. On average, agents produced dozens of socially meaningful interactions per simulated hour—such as conversations, greetings, requests, or confrontations—interspersed with shorter micro-actions (e.g., moving, eating, sleeping) that maintained the flow of daily life.

We extended the original implementation to support controlled experimental manipulations. Specifically, the extension allows features of the environment or of agents, such as group membership, to be specified in natural language and made perceptible to other agents during interaction, while threat stimuli are delivered as private inputs to each agent’s own perception. Agents can therefore perceive stable attributes of others and shared environmental cues, but they only infer others’ perceived threat indirectly through communication. Agents’ memories, plans, and dialogues were updated continuously over three simulated days, producing rich, unscripted trajectories of decisions, movements, and conversations. In addition, we implemented a runtime probing module—analogous to ecological momentary assessment [EMA; 53, 54]—to periodically sample agents’ attitudinal states without affecting behavior (see details in Section 4.4).

Simulations were implemented using the `Mistral-Small` language model ([mistralai/Mistral-Small-24B-Instruct-2501¹](https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501)), accessed via HuggingFace. To improve inference efficiency and reduce GPU memory requirements, we used a quantized version of the model ([matatonic/Mistral-Small-24B-Instruct-2501-6.5bpw-h8-ex12²](https://huggingface.co/matatonic/Mistral-Small-24B-Instruct-2501-6.5bpw-h8-ex12)) executed with the `ExLlamaV2` (v0.2.3) inference engine³. Quantization represents model weights with reduced numerical precision—here, an average of 6.5 bits per weight—while maintaining comparable performance to the full-precision model. This approach substantially decreases memory usage (approximately 60% reduction in VRAM) and accelerates inference with minimal degradation in generative quality. The applied quantization method follows the principles of GPTQ [55], which performs accurate post-training quantization for large generative transformers. Quantized models have been shown to yield large computational and memory gains with negligible impact on output quality across a range of architectures [56].

All simulations were run on an on-demand computing cluster equipped with 40 NVIDIA RTX 4090 GPUs (24 GB VRAM each). The main simulation set was executed in parallel across all GPUs over 2.5 days (approximately 2,400 GPU hours), while the structural condition analyses required 7.5 days (approximately 7,200 GPU hours).

¹[www.huggingface.com/mistralai/Mistral-Small-24B-Instruct-2501](https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501)

²[www.huggingface.com/matatonic/Mistral-Small-24B-Instruct-2501-6.5bpw-h8-ex12](https://huggingface.co/matatonic/Mistral-Small-24B-Instruct-2501-6.5bpw-h8-ex12)

³<https://github.com/turboderp-org/exllamav2>

The choice of the Mistral architecture was motivated by its relatively low alignment and instruction-guarding constraints compared to larger instruction-tuned models such as GPT-4o or Claude 3.7. Preliminary testing with those systems led to full suppression of hostility or refusals to produce any negative intergroup behaviors. Mistral provided a more neutral generative prior, enabling the expression of both cooperative and antagonistic (including hateful) responses essential for modeling realistic social conflict dynamics.

All code for simulation orchestration, quantized inference, and data logging was implemented in Python (v3.10).

4.3 Experimental design and manipulations

For the first set of simulations, we implemented a 2×2 factorial design crossing *realistic threat* (strong vs. none) and *symbolic threat* (strong vs. none). Agents were assigned to one of two minimal groups using natural-language identity prompts (“<Name> is a member of Group A. There is another group, Group B, which they are not part of.”). To account for stochasticity, each condition was run 10 times, for a total of 40 simulations. In each run, we used a different random seed for LLM generation and independently randomized agent-level variables (e.g., group membership). Because agents repeatedly decide where to go, what to do, and what to say, with each decision shaping subsequent choices and others’ reactions, these stochastic elements compound over time and produce distinct trajectories of interaction in each run. All analyses use data from all 10 runs per condition.

Threat manipulations were induced by injecting standardized statements into agents’ perceptions (i.e., what was used as context in agents decision-making prompts) and memories. Statements were adapted from established threat-scale items [57] (e.g., for realistic threat: “<Name> feels that the physical safety of <Group 1> members is threatened by <Group 2>.”; for symbolic threat: “<Name> feels that the values and traditions of <Group 1> are threatened by <Group 2>.”; see Supplementary Table A5). These belief percepts were continuously embedded within agents’ perception and memory streams, ensuring that the intended threat information remained salient and accessible during planning, interaction, and reflection (i.e., continuously added to the agents’ context window when planning, acting, conversing). In effect, the manipulation maintained a stable, high-intensity representation of the target threat type while other threat dimensions were held at baseline.

To examine structural moderators, we ran a second set of simulations that introduced (i) spatial segregation, implemented by assigning agents to groups based on k -means clustering of their home coordinates on the town map, forming geographically distinct clusters that minimized intergroup overlap and maximized mean distance between groups; and (ii) demographic asymmetry—unequal group sizes implemented by varying relative group size, assigning 80% of agents to form the majority group and the remaining 20% to form the minority group.

These manipulations altered the structural configuration of the environment rather than agents’ internal beliefs, and were not reflected in any instructions or prompts. Agents were not told whether they belonged to a majority or minority group, or whether groups were segregated—they simply experienced these conditions through

their environment. This yielded a fully crossed 2 (symbolic threat) \times 2 (realistic threat) \times 2 (segregated–integrated) \times 2 (equal–unequal group size) design, comprising 160 additional simulations. The underlying agent architecture and environment were otherwise identical across all conditions.

Each simulation ran for three simulated days, allowing multi-day interaction and cumulative social experience to shape conflict dynamics, while maintaining computational feasibility⁴.

Each environment contained the same twenty-five agents from Park et al. [10], whose personas represent realistic variation in age (19–68), gender (44% female), occupation (e.g., artist, shopkeeper, student, engineer), personality, and social life to emulate plausible social heterogeneity in everyday settings of a small community. For group assignment, agents were randomly split into two minimal groups (12 vs. 13 members), and in the structural simulations the majority–minority condition used a fixed 20 vs. 5 split.

Finally, we report additional replication sets varying agents’ moral-value profiles and non-minimal group paradigms (pilot) in the SI (Sets 0 and 3; SI Section E.3).

4.4 Agent probing and attitudinal measures

During runtime, agents were periodically probed via separate natural-language queries—analogous to ecological momentary assessment (EMA) in human research. Probes were administered in parallel to agents’ ongoing decision and interaction processes by temporarily copying their current state (i.e., the information and memory at the time of action or conversation) and eliciting scale responses from this duplicate context. This ensured that the probes captured attitudinal states contemporaneous with decision-making while leaving the primary simulation trajectory unaffected. These probes adapted validated items from social–psychological scales measuring intergroup trust [58], collaboration [59], dehumanization [60], and ingroup identification (adapted from group identity, identity fusion, and group commitment scales; [61–64]). Items were phrased in first-person form (e.g., “I consider <Group 1> members to be honest and reliable”) and scored using the model’s numerical responses to 7-point Likert-type anchors (1=totally disagree, 4=neutral, 7=totally agree). Probing occurred continuously throughout the simulation—following new actions or conversations (probing $N = 46,240$)—yielding dynamic, time-resolved measures of trust, cooperation, ingroup bias, identification, and dehumanization attitudes. Full item lists, and reliability statistics are provided in the SI (Section A.2).

4.5 Data logging and derived variables

Across the base and structural simulation sets, the agents produced a rich corpus of social behavior. In the baseline forty simulations (without structural manipulations), 527,387 actions were recorded, including 473,295 agent–agent interactions (238,463 intergroup) and more than 20,000 conversations over three simulated days. The extended structural set (160 simulations) generated over two million actions,

⁴Extending the simulation by one virtual day would increase runtime and GPU cost by 25–30%: roughly 600 additional GPU hours for the main design and 1,800 GPU hours for structural factors analysis.

including 986,401 agent–agent interactions (226,592 intergroup), and approximately 90,000 conversations (\approx 65–75 million tokens). Statistical models were fit to data aggregated at the hourly level, as individual actions occur at highly variable temporal scales and numerous intermediate microactions (e.g., sleeping, moving, or eating) separate socially meaningful events such as conflict or contact, rendering moment-to-moment analyses unstable. This aggregation yielded between 15,000 and 75,000 samples per model, depending on whether actions, conversations, or attitudes were modeled, which differ in event frequency and sampling resolution.

All agent actions, plans, and conversations were logged with timestamps, acting and target agent IDs, group memberships, and experimental conditions. Conversations were analyzed using pretrained classifiers for hateful language, moral language, and sentiment. Specifically, we applied the `Elron/deberta-v3-large-sentiment`⁵ model for sentiment polarity (F1: 0.74), `Elron/deberta-v3-large-hate`⁶ for hate speech detection (F1: 0.61), and a `roberta-base`[65] model fine-tuned on the Moral Foundations Twitter Corpus [MFTC; 66] to distinguish binding versus individualizing moral language (F1: 0.76). For non-linguistic behavioral data, a separate large language model (`Mistral-Large`⁷) classified whether each logged action was hostile (hateful, violent, aggressive) toward outgroup members (see SI Section D.1.1).

The final dataset is event-level, with one row per logged agent action (and, when applicable, per conversation). Each row contains the timestamp; initiator and target IDs (and their group memberships); the run’s condition/structural settings; a brief action log; and a hostility flag. Conversation rows additionally include derived linguistic features (sentiment, hate, moral language). From initiator/target groups we derive intergroup status and non-hostile intergroup contact (intergroup with hostility = false), and attitudinal probe responses are recorded at their sampling times.

4.6 Analytical strategy

Analyses focused on effects of threat type and structural variables on hostile action rates, language content (hateful language, moral language, sentiment), and attitudes over time. The primary specification was a mixed-effects negative binomial regression predicting the hourly rate of hostile actions from realistic and symbolic threat perception, their interaction, non-hostile intergroup contact rate in the previous hour, the hostile action rate in the previous hour (autoregression), a log offset for total actions, and time, with random intercepts for agents and simulation runs (nested within condition). All models were estimated in R (v4.3) using `glmmTMB`. Analogously, additional mixed-effects models with the same fixed-effect structure and random intercepts for agents and runs were fit for language hatefulness, moralization, sentiment, and attitudinal outcomes. System-level analyses aggregated data across all agents within each virtual town and run to characterize collective patterns in behavior, language, and attitudes. Full model results are provided in the Supplementary Information. All code files are available in the project repository <https://osf.io/5ac3d>.

⁵<https://huggingface.co/Elron/deberta-v3-large-emotion>

⁶<https://huggingface.co/Elron/deberta-v3-large-hate>

⁷<https://huggingface.co/mistralai/Mistral-Large-Instruct-2407>

To assess whether attitudes or hateful language mediate the effects of perceived threat on hostile actions, we estimated Bayesian multilevel negative binomial mediation models. Ingroup bias (or hateful language) at time $t-1$ was treated as the mediator linking symbolic and realistic threat perception at time $t-2$ to hostile actions at time t , controlling for prior ingroup bias (or hateful language), intergroup contact rate, and hostile action rate. Models included random intercepts for agents and simulation runs and were estimated on $N = 23,355$ observations. We used weakly informative priors on all parameters: Student- t priors on intercepts, Normal priors on regression coefficients, and Exponential priors on variance and shape parameters. Models were fit in `brms` (via `cmdstanr`) using 4 chains with 3,000 iterations each (1,800 post-warm-up), a 1,200-iteration warm-up, and a target acceptance rate of 0.99. All Bayesian models converged satisfactorily ($\hat{R} < 1.01$, no divergent transitions, and effective sample sizes $> 1,000$ for all parameters). Full Bayesian model outputs are provided in the SI (Supplementary Table B16, B21).

5 Reproducibility, transparency, and robustness

Given the inherently stochastic nature of LLMs and the sequential, interdependent design of the present simulations, reproducibility and transparency are particularly important. Unlike typical one-off LLM uses (classification or single responses), our framework generates chains of decisions in which each agent's plan, action, and dialogue shape later prompts and perceptions, amplifying stochastic variation over time. We therefore provide extensive details on data-generation procedures, model settings, validation, and robustness checks following Abdurahman et al. [67]; see SI Section F.

Data Availability

All simulation data used in this study are available in the project repository at <https://osf.io/5ac3d>. No human data or external datasets were used.

Code Availability

All code for data generation, preprocessing, analysis, and simulations is available in the project repository at <https://osf.io/5ac3d>.

Acknowledgements

This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant No. A9550-23-1-0463. The authors gratefully acknowledge this support.

Author Contributions

Using the CRediT taxonomy, contributions were as follows:

SA: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization.

FKM: Conceptualization, Writing – original draft.

CY: Conceptualization, Methodology, Software, Formal analysis, Writing – review & editing, Visualization.

NK: Conceptualization, Writing – review & editing.

MD: Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

Competing interests

The authors declare no competing interests.

References

- [1] Stephan, W. G. & Stephan, C. W. in *Intergroup threat theory* (eds Macrae, C. N., Hewstone, M. & Hogg, M. A.) *The Oxford Handbook of Intergroup Conflict* 43–59 (Oxford University Press, 2013).
- [2] Waltz, K. N. *Theory of International Politics* (Addison-Wesley, Reading, MA, 1979).
- [3] Mearsheimer, J. J. *The Tragedy of Great Power Politics* (W. W. Norton & Company, New York, 2001).
- [4] Wendt, A. *Social Theory of International Politics* (Cambridge University Press, Cambridge, 1999).
- [5] Fearon, J. D. & Laitin, D. D. Ethnicity, insurgency, and civil war. *American Political Science Review* **97**, 75–90 (2003).
- [6] Collier, P. & Hoeffer, A. Greed and grievance in civil war. *Oxford Economic Papers* **56**, 563–595 (2004).
- [7] Tilly, C. *The Politics of Collective Violence* (Cambridge University Press, Cambridge, 2003).
- [8] Lamont, M. & Molnár, V. The study of boundaries in the social sciences. *Annual Review of Sociology* **28**, 167–195 (2002).
- [9] Pretus, C. *et al.* Neural and behavioral correlates of sacred values and vulnerability to violent extremism. *Frontiers in psychology* **9**, 2462 (2018).
- [10] Park, J. S. *et al.* ACM (ed.) *Generative agents: Interactive simulacra of human behavior.* (ed. ACM) *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, 1–22 (ACM, 2023).
- [11] Gao, C. *et al.* Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences*

Communications **11**, 1–24 (2024).

- [12] Bruch, E. E. & Atwell, J. Agent-based models in empirical social research. *Sociological Methods & Research* **44**, 186–221 (2015).
- [13] Axelrod, R. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration* (Princeton University Press, Princeton, NJ, 1997).
- [14] Lustick, I. S. Ps-i: A user-friendly agent-based modeling platform for testing theories of political identity and action. *Journal of Artificial Societies and Social Simulation* **5** (2002).
- [15] Cederman, L.-E. Modeling the size of wars: From billiard balls to sandpiles. *American Political Science Review* **97**, 135–150 (2003).
- [16] Park, J. S. *et al.* Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109* (2024).
- [17] Cederman, L.-E., Wimmer, A. & Min, B. Why do ethnic groups rebel? new data and analysis. *World Politics* **62**, 87–119 (2010).
- [18] Kalyvas, S. N. *The Logic of Violence in Civil War* (Cambridge University Press, Cambridge, 2006).
- [19] Toft, M. D. Getting religion? the puzzling case of islam and civil war. *International Security* **31**, 97–131 (2007).
- [20] Ginges, J., Atran, S., Medin, D. & Shikaki, K. Sacred bounds on rational resolution of violent political conflict. *Proceedings of the National Academy of Sciences* **104**, 7357–7360 (2007).
- [21] Bauer, M., Cassar, A., Chytilová, J. & Henrich, J. War’s enduring effects on the development of egalitarian motivations and in-group biases. *Psychological science* **25**, 47–57 (2014).
- [22] Voors, M. J. *et al.* Violent conflict and behavior: a field experiment in burundi. *American economic review* **102**, 941–964 (2012).
- [23] Craig, M. A. & Richeson, J. A. On the precipice of a “majority-minority” america: Perceived status threat from the racial demographic shift affects white americans’ political ideology. *Psychological science* **25**, 1189–1197 (2014).
- [24] Kachanoff, F. J., Taylor, D. M., Caouette, J., Khullar, T. H. & Wohl, M. J. The chains on all my people are the chains on me: Restrictions to collective autonomy undermine the personal autonomy and psychological well-being of group members. *Journal of Personality and Social Psychology* **116**, 141 (2019).

- [25] Paolini, S., Harwood, J. & Rubin, M. Negative intergroup contact makes group memberships salient: Explaining why intergroup conflict endures. *Personality and Social Psychology Bulletin* **36**, 1723–1738 (2010).
- [26] Árnadóttir, K. *et al.* How positive and negative intergroup contact jointly inform minority support for social change: The role of system-fairness beliefs. *British Journal of Social Psychology* **63**, 811–838 (2024).
- [27] Gidron, N. & Hall, P. A. The politics of social status: Economic and cultural roots of the populist right. *The British journal of sociology* **68**, S57–S84 (2017).
- [28] Mudde, C. & Rovira Kaltwasser, C. Studying populism in comparative perspective: Reflections on the contemporary and future research agenda. *Comparative political studies* **51**, 1667–1693 (2018).
- [29] Hopkins, D. J. Politicized places: Explaining where and when immigrants provoke local opposition. *American political science review* **104**, 40–60 (2010).
- [30] Inglehart, R. F. & Norris, P. Trump, brexit, and the rise of populism: Economic have-nots and cultural backlash. *HKS Working Paper Series* (2016).
- [31] Cramer, K. J. *The politics of resentment: Rural consciousness in Wisconsin and the rise of Scott Walker* (University of Chicago Press, 2022).
- [32] Mutz, D. C. Status threat, not economic hardship, explains the 2016 presidential vote. *Proceedings of the National Academy of Sciences* **115**, E4330–E4339 (2018).
- [33] Chen, R., Ardit, A., Sleight, H., Evans, O. & Lindsey, J. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509* (2025).
- [34] Abdurahman, S. *et al.* Perils and opportunities in using large language models in psychological research. *PNAS nexus* **3**, pgae245 (2024).
- [35] Atari, M., Xue, M. J., Park, P. S., Blasi, D. & Henrich, J. Which humans? (2023).
- [36] Raganato, A. & Tiedemann, J. ACL (ed.) *An analysis of encoder representations in transformer-based machine translation.* (ed.ACL) *Proceedings of the 2018 EMNLP workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, 287–297 (2018).
- [37] Jin, M. *et al.* Rambow, O. *et al.* (eds) *Exploring concept depth: How large language models acquire knowledge and concept at different layers?* (eds Rambow, O. *et al.*) *Proceedings of the 31st International Conference on Computational Linguistics*, 558–573 (2025).
- [38] Stephan, W. G. *et al.* The role of threats in the racial attitudes of blacks and whites. *Personality and Social Psychology Bulletin* **28**, 1242–1254 (2002).

- [39] Riek, B. M., Mania, E. W. & Gaertner, S. L. Intergroup threat and outgroup attitudes: A meta-analytic review. *Personality and social psychology review* **10**, 336–353 (2006).
- [40] Stephan, W. G., Ybarra, O., Martnez, C. M., Schwarzwald, J. & Tur-Kaspa, M. Prejudice toward immigrants to spain and israel: An integrated threat theory analysis. *Journal of Cross-Cultural Psychology* **29**, 559–576 (1998).
- [41] Velasco González, K., Verkuyten, M., Weesie, J. & Poppe, E. Prejudice towards muslims in the netherlands: Testing integrated threat theory. *British journal of social psychology* **47**, 667–685 (2008).
- [42] Pettigrew, T. F. & Tropp, L. R. A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology* **90**, 751–783 (2006).
- [43] Reimer, N. K. *et al.* Intergroup contact and social change: Implications of negative and positive contact for collective action in advantaged and disadvantaged groups. *Personality and Social Psychology Bulletin* **43**, 121–136 (2017).
- [44] Strachan, J. W. A. *et al.* Testing theory of mind in large language models and humans. *Nature Human Behaviour* **8**, 1285–1295 (2024).
- [45] Xi, Z. *et al.* The rise and potential of large language model based agents: A survey. *Science China Information Sciences* **68**, 121101 (2025).
- [46] Zhang, J. *et al.* Ku, L.-W., Martins, A. & Srikumar, V. (eds) *Exploring collaboration mechanisms for llm agents: A social psychology view*. (eds Ku, L.-W., Martins, A. & Srikumar, V.) *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14544–14607 (2024).
- [47] Jiang, A. Q. *et al.* Mistral 7b (2023). URL <https://arxiv.org/abs/2310.06825>. [arXiv:2310.06825](https://arxiv.org/abs/2310.06825).
- [48] Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behavioral and brain sciences* **33**, 61–83 (2010).
- [49] Sourati, Z., Ziabari, A. S. & Dehghani, M. The homogenizing effect of large language models on human expression and thought. *arXiv preprint arXiv:2508.01491* (2025).
- [50] Adornetto, C. *et al.* Generative agents in agent-based modeling: Overview, validation, and emerging challenges. *IEEE Transactions on Artificial Intelligence* (2025).
- [51] Elhage, N. *et al.* A mathematical framework for transformer circuits. *Transformer Circuits Thread* **1**, 12 (2021).

- [52] Ameisen, E. *et al.* Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread* **6** (2025).
- [53] Shiffman, S. Ecological momentary assessment (ema) in studies of substance use. *Psychological assessment* **21**, 486 (2009).
- [54] Kirchner, T. R. & Shiffman, S. Ecological momentary assessment. *The Wiley-Blackwell Handbook of Addiction Psychopharmacology* 541–565 (2013).
- [55] Frantar, E., Ashkboos, S., Hoefler, T. & Alistarh, D. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323* (2022).
- [56] Lang, J., Guo, Z. & Huang, S. IEEE (ed.) *A comprehensive study on quantization techniques for large language models*. (ed.IEEE) 2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC), 224–231 (IEEE, 2024).
- [57] Kachanoff, F. J., Bigman, Y. E., Kapsakis, K. & Gray, K. Measuring realistic and symbolic threats of covid-19 and their unique impacts on well-being and adherence to public health behaviors. *Social Psychological and Personality Science* **12**, 603–616 (2021).
- [58] Yamagishi, T. & Yamagishi, M. Trust and commitment in the united states and japan. *Motivation and emotion* **18**, 129–166 (1994).
- [59] Caprara, G. V., Steca, P., Zelli, A. & Capanna, C. A new scale for measuring adults' prosocialness. *European Journal of psychological assessment* **21**, 77–89 (2005).
- [60] Kteily, N., Bruneau, E., Waytz, A. & Cotterill, S. The ascent of man: Theoretical and empirical evidence for blatant dehumanization. *Journal of personality and social psychology* **109**, 901 (2015).
- [61] Leach, C. W. *et al.* Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification. *Journal of personality and social psychology* **95**, 144 (2008).
- [62] Ellemers, N., Spears, R. & Doosje, B. Sticking together or falling apart: In-group identification as a psychological determinant of group commitment versus individual mobility. *Journal of personality and social psychology* **72**, 617 (1997).
- [63] Mael, F. & Ashforth, B. E. Alumni and their alma mater: A partial test of the reformulated model of organizational identification. *Journal of organizational Behavior* **13**, 103–123 (1992).

- [64] Cameron, J. E. A three-factor model of social identity. *Self and identity* **3**, 239–262 (2004).
- [65] Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [66] Hoover, J. *et al.* Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science* **11**, 1057–1071 (2020).
- [67] Abdurahman, S., Salkhordeh Ziabari, A., Moore, A. K., Bartels, D. M. & Dehghani, M. A primer for evaluating large language models in social-science research. *Advances in Methods and Practices in Psychological Science* **8**, 25152459251325174 (2025).
- [68] Cuddy, A. J., Fiske, S. T. & Glick, P. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology* **40**, 61–149 (2008).
- [69] Graham, J. *et al.* in *Moral foundations theory: The pragmatic validity of moral pluralism* (eds Devine, P. & Plant, A.) *Advances in experimental social psychology*, Vol. 47 55–130 (Elsevier, 2013).
- [70] Hosoda, M., Stone-Romero, E. F. & Coats, G. The effects of physical attractiveness on job-related outcomes: A meta-analysis of experimental studies. *Personnel psychology* **56**, 431–462 (2003).
- [71] Spence, J. L., Hornsey, M. J., Stephenson, E. M. & Imuta, K. Is your accent right for the job? a meta-analysis on accent bias in hiring decisions. *Personality and Social Psychology Bulletin* **50**, 371–386 (2024).
- [72] Atari, M. *et al.* Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology* **125**, 1157 (2023).
- [73] He, P., Gao, J. & Chen, W. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing (2021). [arXiv:2111.09543](https://arxiv.org/abs/2111.09543).
- [74] Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L. & Neves, L. Cohn, T., He, Y. & Liu, Y. (eds) *TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification.* (eds Cohn, T., He, Y. & Liu, Y.) *Proceedings of Findings of EMNLP* (2020).

Extended Data

Table 1: Predicting system-level hourly hostile action rate ($M1_{sys}$; $N = 2,050$).

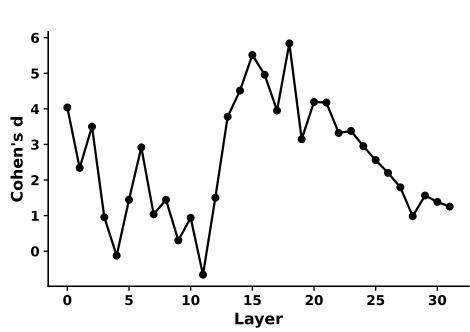
Predictor	β	SE	p
Intercept	-6.96	0.06	< .001
Hostile action rate (lag)	0.16	0.04	< .001
Intergroup contact rate (lag)	-0.30	0.07	< .001
Symbolic threat	0.17	0.05	0.002
Realistic threat	0.33	0.05	< .001
Time	-0.25	0.05	< .001
Symbolic \times Realistic threat	-0.14	0.05	0.007

Table 2: Predicting hourly hateful language rate (M2a; $N = 15,684$).

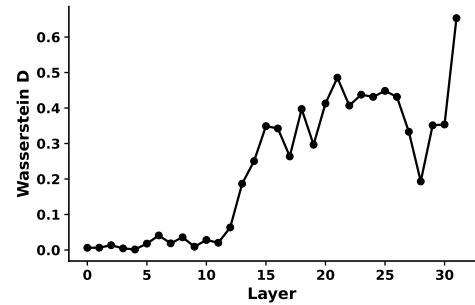
Predictor	β	SE	p
Intercept	-7.27	0.27	< .001
Hateful language rate (lag)	0.03	0.02	0.107
Intergroup contact rate (lag)	0.06	0.05	0.244
Symbolic threat	0.46	0.12	< .001
Realistic threat	0.98	0.12	< .001
Time	-0.01	0.05	0.822
Symbolic \times Realistic threat	-0.28	0.11	0.015

Table 3: Predicting ingroup bias attitudes (M3a; $N = 46,240$).

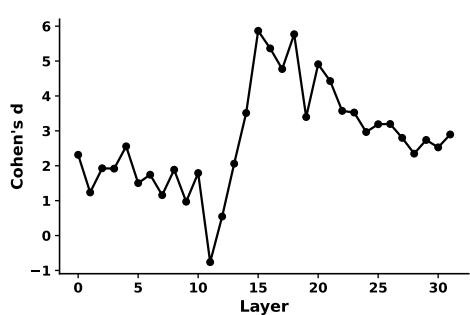
Predictor	β	SE	p
Intercept	0.01	0.06	.88
Group Bias (lag)	0.12	0.01	< .001
Symbolic threat	0.39	0.05	< .001
Realistic threat	0.28	0.05	< .001
Time	-0.01	0.00	.003
Symbolic \times Realistic threat	0.07	0.04	.083



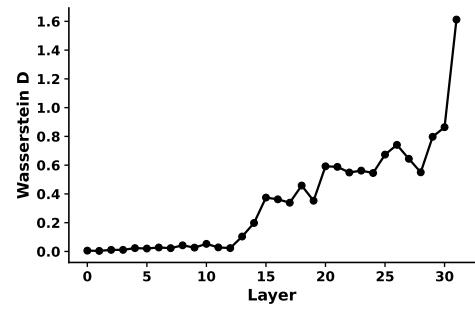
(a) Symbolic vs control: Cohen's d .



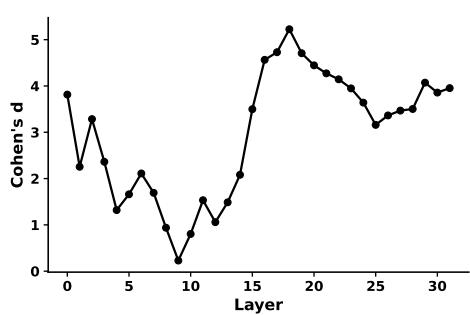
(b) Symbolic vs control: Wasserstein distance.



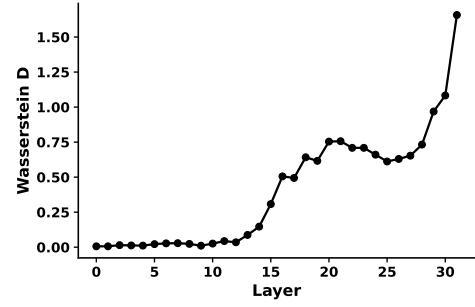
(c) Realistic vs control: Cohen's d .



(d) Realistic vs control: Wasserstein distance.

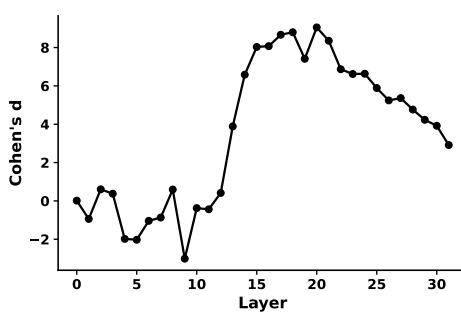


(e) Symbolic vs realistic: Cohen's d .

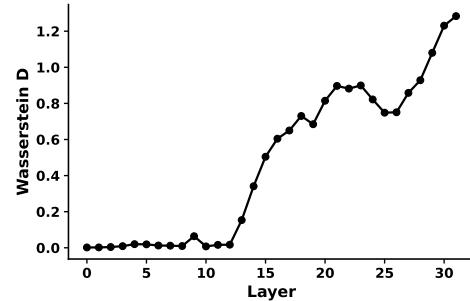


(f) Symbolic vs realistic: Wasserstein distance.

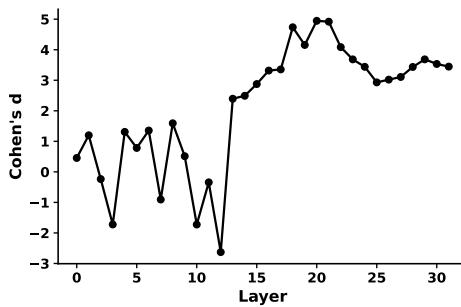
Fig. 1: Layer-wise separability of symbolic and threat realistic threat projections. Panels show Cohen's d and Wasserstein distance across layers for symbolic-threat vs control vignettes (top panels), realistic vs control vignettes (middle panels) and for symbolic-threat vs realistic-threat vignettes (bottom panels).



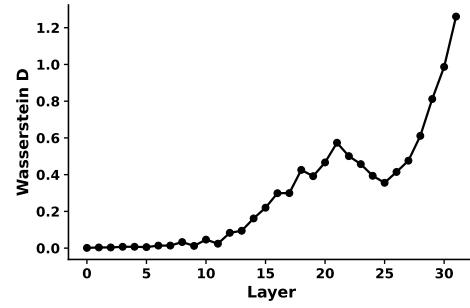
(a) Symbolic vs No-Threat: Cohen's d .



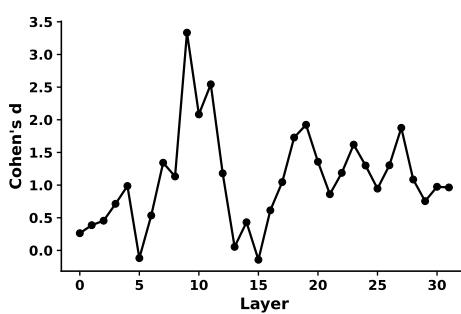
(b) Symbolic vs No-Threat: Wasserstein distance.



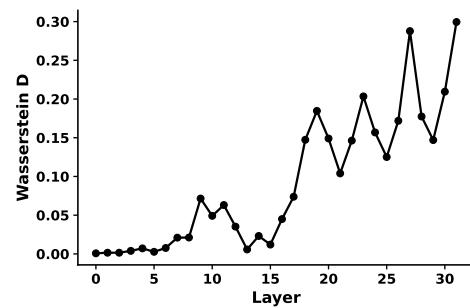
(c) Realistic vs No-Threat: Cohen's d .



(d) Realistic vs No-Threat: Wasserstein distance.



(e) Symbolic vs Realistic: Cohen's d .



(f) Symbolic vs Realistic: Wasserstein distance.

Fig. 2: Layer-wise separability of experimental condition based on their projection on the threat-state vectors. Panels show Cohen's d and Wasserstein distance across layers for symbolic-only vs no-threat condition (top panels), realistic-only vs no-threat condition (middle panels) and for symbolic-only vs realistic-only condition (bottom panels).

Table 4: Predicting hourly hostile action rate including structural manipulations ($M1_{structure}$; $N = 74,487$).

Predictor	β	SE	p
Intercept	-9.83	0.49	< .001
Intergroup contact rate (lag)	-0.58	0.07	< .001
Hostile action rate (lag)	0.05	0.01	< .001
Symbolic threat	0.13	0.05	0.011
Realistic threat	0.28	0.05	< .001
Time	-0.22	0.04	< .001
Segregation	-0.40	0.28	0.152
Group Size	0.08	0.28	0.788
Symbolic \times Realistic	-0.09	0.04	0.041
Segregation \times Group Size	0.42	0.28	0.135

Appendix A Additional Materials and Stimuli

A.1 Threat perception prompts

To induce the intended threat perceptions, we injected natural-language statements into agents' memory and perception streams. These prompts operationalized *realistic* and *symbolic* threat manipulations within a 2×2 factorial design (realistic: strong vs. none; symbolic: strong vs. none). Each condition combined reinforcing statements for the target threat type with suppressing statements for the non-target type, ensuring clear differentiation across experimental cells. Table A5 lists the full set of statements used for each threat condition.

Table A5: Belief statements used to induce or suppress perceived threat.

Threat type	Induce condition	Suppress condition
Symbolic threat	<Name> strongly feels that their values are under threat by <Group 2>.	<Name> does not feel that their values are under threat by <Group 2>.
	<Name> strongly feels that their traditions are under threat by <Group 2>.	<Name> does not feel that their traditions are under threat by <Group 2>.
Realistic threat	<Name> strongly feels economically threatened by <Group 2>.	<Name> does not feel economically threatened by <Group 2>.
	<Name> strongly feels physically threatened by <Group 2>.	<Name> does not feel physically threatened by <Group 2>.

A.2 Scales for Probing Agents

Table A6: Group Identification Items

#	Item
1	I feel a strong bond with <Group 1>.
2	Being a <Group 1> is an important reflection of who I am.
3	I strongly identify with <Group 1>'s values.
4	I feel committed to <Group 1>.
5	I am willing to make personal sacrifices to support <Group 1>.

Note: Adapted from multiple group identity scales [61–64].

Table A7: Group Bias Items

#	Item
1	Members of my group (<Group 1>) are more moral than members of the other group (<Group 2>).
2	Members of my group (<Group 1>) are more trustworthy than members of the other group (<Group 2>).
3	Members of my group (<Group 1>) are more competent than members of the other group (<Group 2>).
4	Members of my group (<Group 1>) are more kind than members of the other group (<Group 2>).

Note: Items adapted from warmth–competence framework questionnaires [68].

Table A8: Trust Items

#	Item
1	I consider <Group 1> members to be honest and reliable.
2	I expect most <Group 1> members to act in good faith.
3	I consider <Group 2> members to be honest and reliable.
4	I expect most <Group 2> members to act in good faith.

Note: Adapted from the Trust and Commitment Scale [58].

Table A9: Collaboration Items

#	Item
1	I am willing to collaborate with <Group 1> members to achieve shared goals.
2	I am happy to share my resources with <Group 1> members when they need help.
3	I am willing to collaborate with <Group 2> members to achieve shared goals.
4	I am happy to share my resources with <Group 2> members when they need help.

Note: Adapted from the Prosocialness Scale [59].

Table A10: Dehumanization Items

#	Item
1	<Group 1> members often seem primitive and uncivilized.
2	<Group 1> members often seem cold and mechanical.
3	<Group 2> members often seem primitive and uncivilized.
4	<Group 2> members often seem cold and mechanical.

Note: Adapted from the Blatant Dehumanization Scale [60].

Appendix B Supplement to Main Analysis

This Appendix reports supplementary and robustness analyses complementing the main text. Unless noted otherwise, count and rate outcomes (e.g., hostile action rate, hateful language rate, moralized language rate) were estimated using mixed-effects negative binomial regressions, and continuous outcomes (e.g., sentiment, attitudes) using Gaussian mixed-effects regressions with random intercepts for agent and simulation run. All models were fit on hourly data as specified in the main text.

B.1 Zero-order correlations of main variables

Table B11 reports zero-order correlations among all key constructs, computed at the agent level by averaging each variable across the three simulated days. These descriptive associations provide an overview of how threat, hostile behavior, language use, attitudes, and intergroup contact co-occur across the simulations.

Table B11: Zero-order correlations among agent-level variables (averaged across the three-day simulation).

	Real.	Symb.	Host.	Hate	Sent.	Bias	Cont.	Bind.	Indiv.
Real. threat	1.00	0.00	0.05	0.25	-0.44	0.53	0.03	0.47	0.82
Symb. threat	0.00	1.00	0.03	0.10	-0.20	0.74	0.08	0.45	0.23
Hostile rate	0.05	0.03	1.00	0.80	-0.72	0.30	0.01	0.26	0.32
Hate lang.	0.25	0.10	0.80	1.00	-0.81	0.47	0.04	0.33	0.49
Sentiment	-0.44	-0.20	-0.72	-0.81	1.00	-0.65	0.02	-0.45	-0.62
Group bias	0.53	0.74	0.30	0.47	-0.65	1.00	0.05	0.59	0.67
Contact	0.03	0.08	0.01	0.04	0.02	0.05	1.00	0.25	0.07
Bind. moral	0.47	0.45	0.26	0.33	-0.45	0.59	0.25	1.00	0.80
Indiv. moral	0.82	0.23	0.32	0.49	-0.62	0.67	0.07	0.80	1.00

B.2 Language dynamics

B.2.1 Hateful language

Model M2a examines whether threat perceptions and prior interaction history predict the rate of hateful language, using hourly conversation-level data.

B.2.2 Sentiment

Model M2b predicts average sentiment within an hour using the same hourly conversation data ($N = 15,684$).

B.2.3 Moral language

Model M2c and Model M2d examine how threat shapes the moral content of conversations, distinguishing binding (loyalty, authority, purity) and individualizing

Table B12: Predicting hourly hateful language rate (M2a; $N = 15,684$).

Predictor	β	SE	p
Intercept	-7.27	0.27	< .001
Hateful language rate (lag)	0.03	0.02	0.107
Intergroup contact rate (lag)	0.06	0.05	0.244
Symbolic threat	0.46	0.12	< .001
Realistic threat	0.98	0.12	< .001
Time	-0.01	0.05	0.822
Symbolic \times Realistic threat	-0.28	0.11	0.015

Table B13: Predicting hourly mean sentiment (M2b).

Predictor	β	SE	p
Intercept	1.70	0.04	< .001
Intergroup contact rate (lag)	0.00	0.00	0.032
Symbolic threat	-0.05	0.01	< .001
Realistic threat	-0.10	0.01	< .001
Time	0.02	0.00	< .001
Sentiment (lag)	0.03	0.00	< .001
Symbolic \times Realistic threat	-0.01	0.01	0.080

(harm/care, fairness) moral language based on Moral Foundations Theory [69]. Both models predict the hourly rate of moralized language use using the same hourly conversation data ($N = 15,684$).

Table B14: (Panel A). Predicting hourly rate of binding moral language (M2c).

Predictor	β	SE	p
Intercept	-1.46	0.05	< .001
Binding (lag)	0.08	0.01	< .001
Intergroup contact rate (lag)	-0.02	0.01	.007
Symbolic threat	0.31	0.02	< .001
Realistic threat	0.33	0.02	< .001
Time	0.03	0.01	< .001
Symbolic \times Realistic threat	-0.16	0.02	< .001

Table B15: (Panel B). Predicting hourly rate of individualizing moral language (M2d).

Predictor	β	SE	p
Intercept	-2.62	0.08	< .001
Individualizing (lag)	0.07	0.01	< .001
Intergroup contact rate (lag)	-0.01	0.01	.203
Symbolic threat	0.08	0.06	.157
Realistic threat	0.64	0.06	< .001
Time	0.04	0.01	< .001
Symbolic \times Realistic threat	0.03	0.06	.614

B.2.4 Mediation of threat effects on hostile action rate through hateful language

To assess whether hateful language mediates the effects of perceived threat on hostile actions, we estimated a Bayesian multilevel mediation model (**brms**). Hateful language rate at time $t - 1$ was specified as the mediator linking symbolic and realistic threat perception (at $t - 2$) to hostile actions at time t , controlling for prior hateful language, intergroup contact rate, and hostile action rate, with random intercepts for agents and simulation run ($N = 23,355$). We used weakly informative priors on all parameters (Student-t priors on intercepts, Normal priors on regression coefficients, and exponential priors on variance and shape parameters).

Posterior estimates indicated small but credibly positive a paths from both threat types to hateful language ($\hat{\beta}_{a,\text{realistic}} = 0.06$ [0.04, 0.07]; $\hat{\beta}_{a,\text{symbolic}} = 0.03$ [0.01, 0.04]). By contrast, the b path from hateful language to hostile actions was slightly negative and practically small ($\hat{\beta}_b = -0.05$ [-0.11, 0.00]), indicating that higher levels of hateful language do not predict increases in subsequent hostile actions and may, if anything, be weakly attenuating. Consistent with this, the resulting indirect effects were extremely small, negative, and effectively centered on zero ($\hat{\beta}_{\text{indirect, realistic}} = -0.00$ [-0.01, 0.00]; $\hat{\beta}_{\text{indirect, symbolic}} = -0.00$ [0.00, 0.00]), providing no evidence that hateful language mediates the effects of threat on hostility.

In contrast, the direct effect of realistic threat on hostile actions remained robustly positive ($\hat{\beta}_{\text{direct, realistic}} = 0.26$ [0.10, 0.43]), while the direct effect of symbolic threat was positive but more uncertain ($\hat{\beta}_{\text{direct, symbolic}} = 0.13$ [-0.04, 0.30]). Overall, these results indicate that (1) hateful language does not escalate hostile behavior, (2) hateful language does not constitute a meaningful mediating pathway from threat to hostility, and (3) hostile actions are driven primarily by perceived threat, especially realistic threat, via a direct route rather than via hateful language.

Table B16: Bayesian mediation of threat effects on hostile actions via hateful language ($N = 23,355$).

Effect	Estimate	95% CrI	pd (%)
a (Realistic threat \rightarrow Hateful language)	0.06	[0.04, 0.07]	100.00
a (Symbolic threat \rightarrow Hateful language)	0.03	[0.01, 0.04]	99.36
b (Hateful language \rightarrow Hostility)	-0.05	[-0.11, 0.00]	98.09
Indirect (Realistic threat \rightarrow Hateful language \rightarrow Hostility)	-0.00	[-0.01, 0.00]	98.09
Indirect (Symbolic threat \rightarrow Hateful language \rightarrow Hostility)	-0.00	[0.00, 0.00]	97.45
Direct (Realistic threat \rightarrow Hostility)	0.26	[0.10, 0.43]	99.90
Direct (Symbolic threat \rightarrow Hostility)	0.13	[-0.04, 0.30]	93.67

All estimates are posterior medians with 95% credible intervals (CrI) and posterior probability of direction (pd)

B.3 Attitudes

B.3.1 Ingroup bias attitudes

Model M3a predicts agents' average ingroup bias scores using the data from probing agents throughout the simulation ($N = 46,240$). Agents were probed when making a decision by eliciting responses to group identity scales.

Table B17: Predicting ingroup bias attitudes (M3a).

Predictor	β	SE	p
Intercept	0.01	0.06	.88
Group Bias (lag)	0.12	0.01	< .001
Symbolic threat	0.39	0.05	< .001
Realistic threat	0.28	0.05	< .001
Time	-0.01	0.00	0.003
Symbolic \times Realistic threat	0.07	0.04	0.083

B.3.2 Ingroup identity attitudes

Model M3b predicts agents' average group identity scores using the data from probing agents throughout the simulation ($N = 46,240$). Agents were probed when making a decision by eliciting responses to group identity scales.

B.3.3 Predicting attitudes from prior hateful language

We tested whether hateful language predicted subsequent changes in ingroup bias, controlling for prior non-hostile intergroup contact, prior hostile action rate, prior attitudes, and threat conditions (Table B19; $N = 37,105$). This analysis evaluated whether

Table B18: Predicting group identity attitudes (M3b).

Predictor	β	SE	<i>p</i>
Intercept	0.00	0.01	.89
Group Identity (lag)	0.64	0.01	< .001
Symbolic threat	0.24	0.01	< .001
Realistic threat	0.07	0.01	< .001
Time	0.10	0.00	< .001
Symbolic \times Realistic threat	-0.01	0.01	0.083

language functions as a driver of cognitive orientations. We found no evidence that hateful language predicted later attitudes, whereas prior hostile action rate and threat did. This suggests that language does not shape attitudes but rather co-occurs with threat-induced ingroup bias. Note that hostile action rate did predict more ingroup bias (0.02, *p* = .002; Table B19), indicating potential reinforcement effects and feedback loops of ingroup bias increasing hostile action rate which then make agents more biased against each other leading to an more hostile actions.

Table B19: Predicting bias attitudes from prior hateful language.

Predictor	β	SE	<i>p</i>
Intercept	4.43	0.07	< .001
Intergroup contact rate (lag)	-0.00	0.01	0.563
Hostile action rate (lag)	0.02	0.01	0.002
Hateful language rate (lag)	0.01	0.01	0.363
Group Bias (lag)	0.15	0.01	< .001
Symbolic threat	0.43	0.06	< .001
Realistic threat	0.31	0.06	< .001
Time	-0.02	0.01	0.001
Symbolic \times Realistic threat	0.10	0.06	0.086

B.3.4 Predicting language from prior group bias

We next tested the reverse relationship, whether prior group bias predicted later hateful language, controlling for prior intergroup contact rate, hostile action rate, and threat manipulations (Table B20; *N* = 37,105). This model tested whether attitudes contribute to subsequent linguistic hostility. We do not find evidence that attitudes predict later hateful language when controlling for threat and prior hostility. Hateful language therefore appears as a concurrent expression of perceived threat, not a downstream product of group bias.

Table B20: Predicting hateful language from prior attitudes.

Predictor	β	SE	p
Intercept	-5.01	0.26	< .001
Intergroup contact rate (lag)	0.04	0.06	0.478
Hostile action rate (lag)	0.03	0.02	0.078
Hateful language rate (lag)	0.03	0.02	0.258
Group Bias (lag)	0.10	0.06	0.130
Symbolic threat	0.47	0.11	< .001
Realistic threat	0.81	0.11	< .001
Time	-0.06	0.06	0.267
Symbolic \times Realistic threat	-0.27	0.11	0.014

B.3.5 Mediation of threat effects on hostile action rate through group bias

To assess whether attitudinal bias mediates the effects of perceived threat on hostile actions, we estimated a Bayesian multilevel mediation model (`brms`). Group bias at time $t-1$ was specified as the mediator linking symbolic and realistic threat perception (at $t-2$) to hostile actions at time t , controlling for prior group bias, intergroup contact rate, and hostile action rate, with random intercepts for agents and simulation run ($N = 23,355$). We used weakly informative priors on all parameters (Student-t priors on intercepts, Normal priors on regression coefficients, and exponential priors on variance and shape parameters).

Posterior estimates showed sizable positive a paths from both threat types to group bias, mirroring the main attitudes model ($\hat{\beta}_{a,\text{realistic}} = 0.28$ [0.16, 0.39]; $\hat{\beta}_{a,\text{symbolic}} = 0.38$ [0.27, 0.50]). In turn, higher bias at time $t-1$ predicted more hostile actions at time t ($\hat{\beta}_b = 0.25$ [0.07, 0.43]). As a result, we observed credible indirect effects of both realistic and symbolic threat on hostility via bias ($\hat{\beta}_{\text{indirect, realistic}} = 0.07$ [0.02, 0.14]; $\hat{\beta}_{\text{indirect, symbolic}} = 0.09$ [0.02, 0.18]). The direct effect of realistic threat remained positive ($\hat{\beta}_{\text{direct, realistic}} = 0.17$ [0.02, 0.37]), whereas symbolic threat showed essentially no direct effect on hostile actions ($\hat{\beta}_{\text{direct, symbolic}} = -0.00$ [-0.21, 0.21]).

Taken together, these results indicate that both realistic and symbolic threat reliably increase group bias, and that this elevated bias, in turn, predicts more hostile actions. Hostile behavior thus appears to be substantially mediated by intergroup bias for both threat types, with realistic threat additionally exerting a sizable residual direct effect on hostility.

Table B21: Bayesian mediation of threat effects on hostile actions via ingroup bias ($N = 23,355$).

Effect	Estimate	95% CrI	pd (%)
a (Realistic threat \rightarrow Bias)	0.28	[0.16, 0.39]	100.00
a (Symbolic threat \rightarrow Bias)	0.38	[0.27, 0.50]	100.00
b (Bias \rightarrow Hostility)	0.25	[0.07, 0.43]	99.65
Indirect (Realistic threat \rightarrow Bias \rightarrow Hostility)	0.07	[0.02, 0.14]	99.65
Indirect (Symbolic threat \rightarrow Bias \rightarrow Hostility)	0.09	[0.02, 0.18]	99.65
Direct (Realistic threat \rightarrow Hostility)	0.17	[0.02, 0.37]	95.71
Direct (Symbolic threat \rightarrow Hostility)	-0.00	[-0.21, 0.21]	50.65

All estimates are posterior medians with 95% credible intervals (CrI) and posterior probability of direction (pd)

B.4 System-level Models

To assess whether the patterns observed at the agent level were also present at the collective level, we repeated the main hostile action (M1) and hateful language (M2a) models using system-level hourly counts, aggregated across all agents. Model structure was otherwise identical. This robustness check evaluates whether aggregate group dynamics (e.g., total occurrence of hostile actions in the town) mirror agent-level processes.

System-level models yielded near-identical patterns to agent-level analyses: realistic threat perception exerted the strongest effects, symbolic threat perception had a smaller effect, and the interaction between threat types was negative. Prior hostile action rate showed somewhat stronger autocorrelation at the system-level, reflecting the persistence of collective hostility once it emerged. Overall, these findings confirm that the dynamics observed at the individual level scale up to the collective level.

Table B22: Predicting system-level hourly hostile action rate ($N = 2,050$).

Predictor	β	SE	p
Intercept	-6.96	0.06	< .001
Hostile action rate (lag)	0.16	0.04	< .001
Intergroup contact rate (lag)	-0.30	0.07	< .001
Symbolic threat	0.17	0.05	0.002
Realistic threat	0.33	0.05	< .001
Time	-0.25	0.05	< .001
Symbolic \times Realistic threat	-0.14	0.05	0.007

Table B23: Predicting system-level hourly hateful language rates ($N = 1,654$).

Predictor	β	SE	p
Intercept	-6.36	0.12	< .001
Hateful language rate (lag)	0.03	0.04	0.542
Intergroup contact rate (lag)	-0.05	0.07	0.426
Symbolic threat	0.44	0.11	< .001
Realistic threat	0.98	0.11	< .001
Time	-0.04	0.05	0.460
Symbolic \times Realistic threat	-0.25	0.11	0.022

Appendix C Supplement to Structural Contexts Analyses

Unless noted otherwise, count and rate outcomes (e.g., hostile action rate, hateful language rate, moralized language rate) were estimated using mixed-effects negative binomial regressions, and continuous outcomes (e.g., sentiment, attitudes) using Gaussian mixed-effects regressions with random intercepts for agents and simulation run. All models were fit on hourly data as specified in the main text.

C.1 Hostile actions models

Table C24: Predicting hourly hostile action rate ($M1_{structure}$; $N = 74,487$).

Predictor	β	SE	p
Intercept	-9.83	0.49	< .001
Intergroup contact rate (lag)	-0.58	0.07	< .001
Hostile action rate (lag)	0.05	0.01	< .001
Symbolic threat	0.13	0.05	0.011
Realistic threat	0.28	0.05	< .001
Time	-0.22	0.04	< .001
Segregation	-0.40	0.28	0.152
Group Size	0.08	0.28	0.788
Symbolic \times Realistic	-0.09	0.04	0.041
Segregation \times Group Size	0.42	0.28	0.135

C.2 Language models

Table C25: Predicting hateful language ($M2a_{structure}$; $N = 32,059$).

Predictor	β	SE	p
Intercept	-7.46	0.28	< .001
Hateful language rate (lag)	0.09	0.02	< .001
Intergroup contact rate (lag)	0.09	0.04	0.028
Symbolic threat	0.42	0.13	0.002
Realistic threat	0.89	0.14	< .001
Time	-0.07	0.04	0.054
Segregation	-0.05	0.09	0.554
Group Size	-0.12	0.09	0.187
Symbolic \times Realistic	-0.41	0.13	0.002
Segregation \times Group Size	-0.08	0.09	0.402

Table C26: Predicting sentiment ($M2b_{structure}$) ($N = 32,059$).

Predictor	β	SE	p
Intercept	1.73	0.04	< .001
Intergroup contact rate (lag)	0.00	0.00	0.660
Sentiment (lag)	0.03	0.00	< .001
Symbolic threat	-0.03	0.00	< .001
Realistic threat	-0.05	0.00	< .001
Time	0.02	0.00	< .001
Segregation	0.02	0.00	< .001
Group Size	0.05	0.00	< .001
Symbolic \times Realistic	0.00	0.00	0.298
Segregation \times Group Size	-0.01	0.00	0.002

Table C27: Predicting binding moral language ($M2c_{structure}$) ($N = 32,059$).

Predictor	β	SE	p
Intercept	-1.64	0.06	< .001
Binding language (lag)	0.10	0.00	< .001
Intergroup contact rate (lag)	-0.01	0.00	0.147
Symbolic threat	0.43	0.02	< .001
Realistic threat	0.43	0.02	< .001
Time	-0.01	0.00	0.014
Segregation	-0.05	0.01	< .001
Group Size	-0.12	0.01	< .001
Symbolic \times Realistic	-0.27	0.02	< .001
Segregation \times Group Size	0.02	0.01	0.005

Table C28: Predicting individualizing moral language ($M2d_{structure}$) ($N = 32,059$).

Predictor	β	SE	p
Intercept	-3.22	0.09	< .001
Individualizing language (lag)	0.12	0.01	< .001
Intergroup contact rate (lag)	-0.01	0.01	0.505
Symbolic threat	0.34	0.05	< .001
Realistic threat	0.94	0.05	< .001
Time	-0.02	0.01	0.082
Segregation	-0.05	0.01	< .001
Group Size	-0.34	0.02	< .001
Symbolic \times Realistic	-0.34	0.05	< .001
Segregation \times Group Size	-0.01	0.01	0.381

C.3 Attitudes models

Table C29: Predicting group bias attitudes ($M3a_{structure}$) ($N = 46,156$).

Predictor	β	SE	p
Intercept	0.03	0.04	0.360
Group Bias (lag)	0.17	0.00	< .001
Symbolic threat	0.32	0.01	< .001
Realistic threat	0.23	0.01	< .001
Time	-0.00	0.00	0.227
Segregation	0.00	0.01	0.832
Group Size	-0.05	0.01	< .001
Symbolic \times Realistic	0.15	0.01	< .001
Segregation \times Group Size	0.00	0.01	0.927

Table C30: Predicting group identity attitudes ($M3b_{structure}$; $N = 46,156$).

Predictor	β	SE	p
Intercept	-0.01	0.01	0.704
Group Identity (lag)	0.66	0.00	< .001
Symbolic threat	0.24	0.01	< .001
Realistic threat	0.08	0.01	< .001
Time	0.08	0.00	< .001
Segregation	0.00	0.00	0.394
Group Size	0.00	0.00	0.312
Symbolic \times Realistic	-0.01	0.01	0.075
Segregation \times Group Size	0.00	0.00	0.965

C.4 Predicting intergroup contact rate from threat and structural conditions

We tested whether realistic or symbolic threat manipulations, or structural features of the simulated environment, predicted the rate of non-hostile intergroup contact events. We find no significant effects of either threat type, indicating that threat did not systematically alter rates of cross-group interaction (Table C34). By contrast, structural factors exerted moderate to strong effects: segregation and majority status both reduced intergroup contact rate, and their interaction showed that majority agents in segregated settings engaged in the fewest cross-group interactions. These results confirm that non-hostile intergroup contact emerged as an autonomous, self-organizing process primarily constrained by structural context rather than driven by threat.

Table C31: Predicting trust attitudes ($M3c_{structure}$) ($N = 17,202$).

Predictor	β	SE	p
Intercept	-0.03	0.02	0.098
Trust (lag)	0.63	0.01	< .001
Symbolic threat	-0.14	0.01	< .001
Realistic threat	-0.14	0.01	< .001
Time	-0.04	0.00	< .001
Segregation	0.00	0.01	0.424
Group Size	-0.04	0.01	< .001
Symbolic \times Realistic	0.11	0.01	< .001
Segregation \times Group Size	0.01	0.01	0.342

Table C32: Predicting collaboration attitudes ($M3d_{structure}$) ($N = 17,202$).

Predictor	β	SE	p
Intercept	-0.03	0.02	0.118
Collaboration (lag)	0.55	0.01	< .001
Symbolic threat	-0.15	0.01	< .001
Realistic threat	-0.15	0.01	< .001
Time	-0.03	0.00	< .001
Segregation	0.01	0.01	0.392
Group Size	-0.06	0.01	< .001
Symbolic \times Realistic	0.14	0.01	< .001
Segregation \times Group Size	0.01	0.01	0.322

Table C33: Predicting dehumanization attitudes ($M3e_{structure}$) ($N = 17,202$).

Predictor	β	SE	p
Intercept	-0.09	0.02	< .001
Dehumanization (lag)	0.18	0.01	< .001
Symbolic threat	0.21	0.01	< .001
Realistic threat	0.40	0.01	< .001
Time	-0.07	0.01	< .001
Segregation	0.01	0.01	0.398
Group Size	-0.06	0.01	< .001
Symbolic \times Realistic	0.20	0.01	< .001
Segregation \times Group Size	0.00	0.01	0.734

Table C34: Predicting intergroup contact rates ($N = 74,487$).

Predictor	β	SE	p
Intercept	-1.75	0.06	< .001
Intergroup contact rate (lag)	0.46	0.01	< .001
Symbolic threat	0.01	0.01	0.470
Realistic threat	-0.01	0.01	0.527
Time (hours)	-0.05	0.01	< .001
Segregation	-0.23	0.01	< .001
Group Size (majority)	-0.55	0.01	< .001
Symbolic \times Realistic threat	0.00	0.01	0.892
Segregation \times Group Size	-0.13	0.01	< .001

C.5 System-Level Models

C.5.1 Hostile actions

We re-estimated the main model (M1) at the system-level, aggregating hostile actions across all agents (i.e., total occurrence of hostile actions in the town) to produce a global hostile action rate for the whole virtual town (Table C35). The results closely mirrored the agent-level analyses. Realistic threat exerted a positive effect on hateful behavior ($\hat{\beta} = 0.29$, $p < .001$), while symbolic threat was weaker ($\hat{\beta} = 0.19$, $p < .001$), their interaction was negative but non-significant ($\hat{\beta} = -0.07$, $p = .056$). Prior intergroup contact rate again reduced hostility ($\hat{\beta} = -0.50$, $p < .001$), whereas prior hostile action rate predicted continued hostility ($\hat{\beta} = 0.12$, $p < .001$). Structural factors also exhibited systematic effects at the collective level. Segregation substantially reduced overall hostility ($\hat{\beta} = -0.95$, $p < .001$), but majority groups displayed greater hostility overall ($\hat{\beta} = 0.51$, $p = .006$), and their dominance intensified under segregation ($\hat{\beta}_{\text{interaction}} = 1.03$, $p < .001$). Together, these findings confirm that the threat–hostility relationship generalizes from individual to collective scales while revealing that structural asymmetries shape the distribution of hostility across groups.

Table C35: Predicting system-level hostile action rate ($N = 7,769$).

Predictor	β	SE	p
Intercept	-7.82	0.18	< .001
Intergroup contact rate (lag)	-0.50	0.08	< .001
Hostile action rate (lag)	0.12	0.01	< .001
Symbolic threat	0.19	0.03	< .001
Realistic threat	0.29	0.03	< .001
Segregation	-0.95	0.18	< .001
Group Size (majority)	0.51	0.19	0.006
Time	-0.32	0.03	< .001
Symbolic \times Realistic threat	-0.07	0.03	0.056
Segregation \times Group Size	1.03	0.18	< .001

C.5.2 Hateful language

We next examined hateful language aggregated at the system-level. Consistent with agent-level results, both threat types increased the rate of hateful language in the system, with realistic threat perception producing much stronger effects and effects of symbolic threat perception nearly vanishing when realistic threats were perceived. Structural asymmetries were again pronounced: segregation reduced hateful language overall, while majority groups produced substantially more under segregation (Table C36).

Table C36: Predicting system-level rate of hateful language ($N = 5,867$).

Predictor	β	SE	p
Intercept	-6.90	0.13	< .001
Hateful language rate (lag)	0.19	0.03	< .001
Intergroup contact rate (lag)	0.10	0.09	0.246
Segregation	-0.38	0.07	< .001
Group Size (majority)	0.37	0.09	< .001
Symbolic threat	0.46	0.12	< .001
Realistic threat	0.89	0.12	< .001
Time	-0.11	0.04	0.008
Segregation \times Group Size	0.36	0.07	< .001
Symbolic \times Realistic threat	-0.39	0.12	0.001

C.5.3 Sentiment

We next examined sentiment aggregated at the system-level. Consistent with agent-level results, both symbolic and realistic threat perceptions reduced overall sentiment, with realistic threat exerting the stronger negative effect and co-occurrence of both threats amplifying this effect. Structural asymmetries also mirrored the agent-level results: segregation and majority-group status were associated with more positive sentiment (Table C37).

Table C37: Predicting system-level sentiment ($N = 5,867$).

Predictor	β	SE	p
Intercept	2.76	0.00	< .001
Sentiment (lag)	0.04	0.00	< .001
Intergroup contact rate (lag)	0.02	0.00	< .001
Segregation	0.03	0.00	< .001
Group Size (majority)	0.05	0.00	< .001
Symbolic threat	-0.04	0.00	< .001
Realistic threat	-0.07	0.00	< .001
Time	0.02	0.00	< .001
Segregation \times Group Size	-0.03	0.00	< .001
Symbolic \times Realistic threat	-0.01	0.00	0.001

C.5.4 Binding moral language

At the collective level, binding moral language increased sharply under both threat types, with a negative interaction indicating non-additivity under combined threat. Segregation reduced binding language overall, and majority-minority differences converged under segregation (Table C38).

Table C38: Predicting system-level rate of binding moral language ($N = 5,867$).

Predictor	β	SE	p
Intercept	-1.68	0.02	< .001
Binding language (lag)	0.14	0.01	< .001
Intergroup contact rate (lag)	-0.05	0.01	< .001
Segregation	-0.10	0.01	< .001
Group Size (majority)	-0.04	0.01	< .001
Symbolic threat	0.41	0.02	< .001
Realistic threat	0.40	0.02	< .001
Time	-0.01	0.01	0.008
Segregation \times Group Size	0.09	0.01	< .001
Symbolic \times Realistic threat	-0.26	0.02	< .001

C.5.5 Individualizing moral language

Finally, individualizing moral language followed a nearly identical pattern to binding moralization but was dominated by realistic threat with the effect of symbolic threat perception vanishing when realistic threats were perceived. We further observed strong threat effects and reduced individualizing language under segregation. The positive segregation \times Group Size interaction indicated convergence in individualizing language between groups under segregation (Table C39).

Table C39: Predicting system-level rate of individualizing moral language ($N = 5,867$).

Predictor	β	SE	p
Intercept	-3.22	0.05	< .001
Individualizing language (lag)	0.17	0.01	< .001
Intergroup contact rate (lag)	-0.07	0.02	< .001
Segregation	-0.10	0.01	< .001
Group Size (majority)	-0.25	0.02	< .001
Symbolic threat	0.36	0.05	< .001
Realistic threat	0.89	0.05	< .001
Time	-0.03	0.01	0.009
Segregation \times Group Size	0.06	0.01	< .001
Symbolic \times Realistic threat	-0.32	0.05	< .001

Appendix D Supplement to Probing LLM Activations

D.1 Prompts

D.1.1 Prompts for detecting hostile actions

To detect intergroup hostile actions in the simulation logs, we used the following prompt:

Analyze the description of an action between a member of GROUP_A and a member of GROUP_B and determine whether the action is hostile, defined as aggressive, harmful, or hateful behavior (e.g., physical or verbal attacks, threats, harassment, trolling, or discriminating against others).

Here is the action:

[ACTION]

Respond only with "yes" or "no".

D.1.2 Prompts for evaluating hostility in steering experiments

We used the following prompt to determine the extent of hostility expressed in an agent's behavior in the steering experiments.

You are an expert evaluator of intergroup behavior.

Scenario:
{scenario}

Person's behavioral response:
{response}

Task:
Evaluate the hostility level of this behavioral response toward Group B.

Provide ONLY valid JSON:

```
{  
    "rating": <float 1.0-5.0>,  
    "behavior_type": "<friendly|neutral|verbal_attack|social_exclusion|physical_threat|discrimination>",  
    "specific_actions": ["<action1>", "<action2>"],  
    "reasoning": "<brief explanation>",  
    "is_hostile": <true|false>  
}
```

Rating scale:
1.0 = Friendly/supportive
2.0 = Neutral
3.0 = Mildly hostile (avoidance)
4.0 = Moderately hostile (verbal aggression)
5.0 = Extremely hostile (threats/violence)

Respond with JSON only.

D.2 Internal Threat States

D.2.1 Results

Validation of Internal Threat States

Table D40: Mean projection values and standard deviations of control-vignettes on each threat-state vector (layer 32).

Threat-state vector	Control Type	<i>M</i>	<i>SD</i>
Symbolic-threat	Symbolic-control	-1.75	0.59
Symbolic-threat	Unrelated-control	0.14	0.48
Realistic-threat	Realistic-control	-1.43	0.65
Realistic-threat	Unrelated-control	0.93	0.48

Table provides an overview of how much the control scenarios activate the respective threat state and shows low or negative values indicating minimal activation.

Table D41: Projections of threat vignettes onto threat-state vectors (layer 32).

Threat-state vector	Vignette Contrast	df	<i>t</i>	<i>p</i>	Cohen's <i>d</i>	<i>D</i>	<i>p</i> _{<i>D</i>}
Symbolic-threat	Symbolic-threat vs symbolic-control	540.6	15.18	< .001	1.25	0.65	< .001
Realistic-threat	Realistic-threat vs realistic-control	528.8	35.34	< .001	2.90	1.61	< .001
Symbolic-vs-realistic	Symbolic-threat vs realistic-threat	552.6	48.01	< .001	3.95	1.66	< .001
Symbolic-threat	Symbolic-threat vs unrelated-control	497.2	32.33	< .001	2.65	1.06	< .001
Realistic-threat	Realistic-threat vs unrelated-control	557.1	30.82	< .001	2.53	1.08	< .001

We report Welch *t*-test degrees of freedom (df), test statistic (*t*), *p*-value, Cohen's *d*, Wasserstein distance *D* between projection distributions, and the associated *p*-value (*p*_{*D*}). All tests are two-tailed. Table shows significantly stronger activation of the threat state in corresponding threat vs control scenarios.

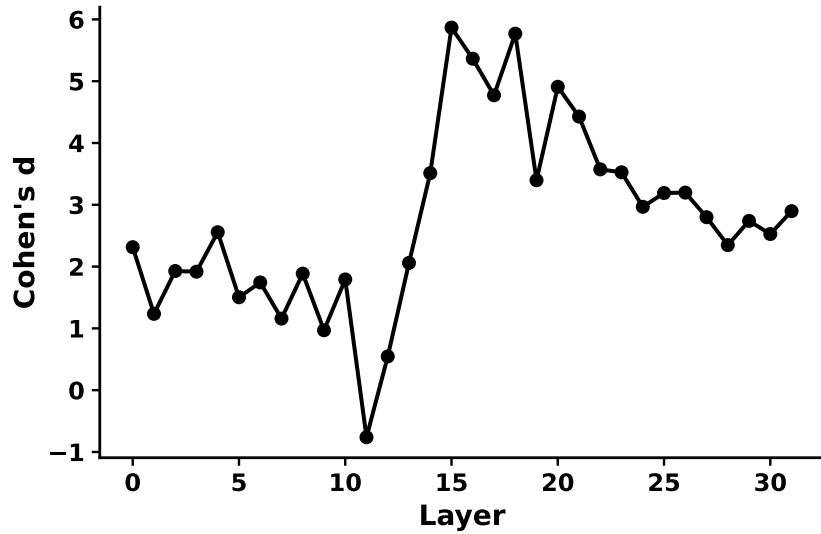


Fig. D3: Difference in projection strength (Cohen's d) between realistic-threat and realistic-control vignettes across layers. Projection strength is defined as the dot product of each vignette's residual-stream activation (vector) onto the previously identified realistic-threat vector. Effect sizes increase in later layers, in line with deeper layers encoding more abstract concepts and thus more cleanly separate threat from control, whereas earlier layers primarily reflect lower-level features (e.g., grammar and structure) that are balanced across vignettes.

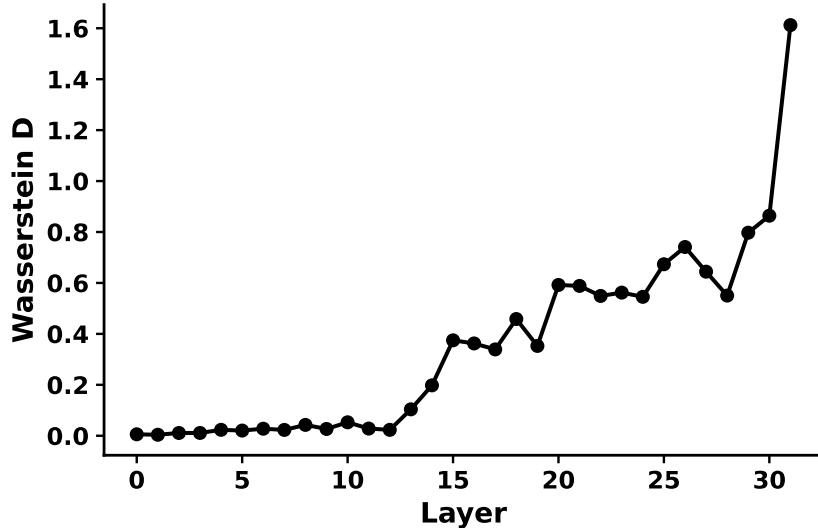


Fig. D4: Wasserstein distance between projection distributions for realistic-threat and realistic-control vignettes across layers. Higher values in later layers indicate strong distributional separation of internal states associated with realistic threat versus no realistic threat.

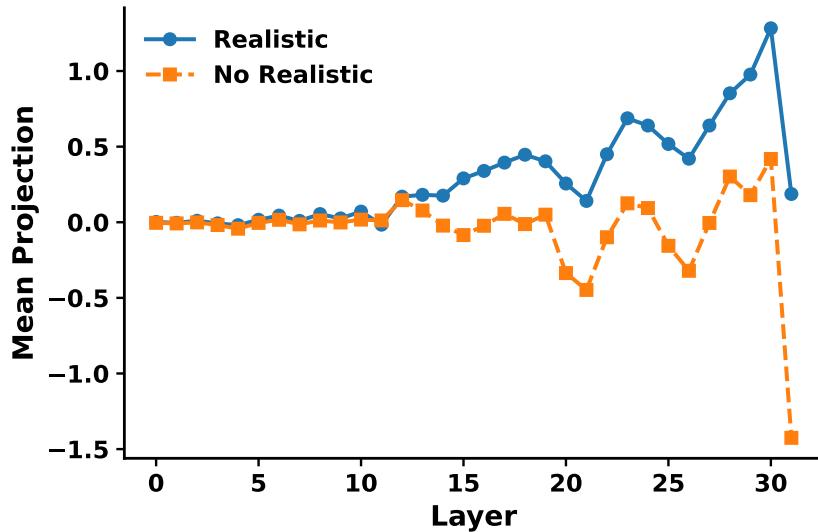


Fig. D5: Mean projection scores for realistic-threat and realistic-control vignettes across layers for the realistic-threat vector. Figure shows that held-out realistic-threat scenarios consistently yield higher projections than control scenarios, which show near zero or negative projections indicating they did not activate threat representations.

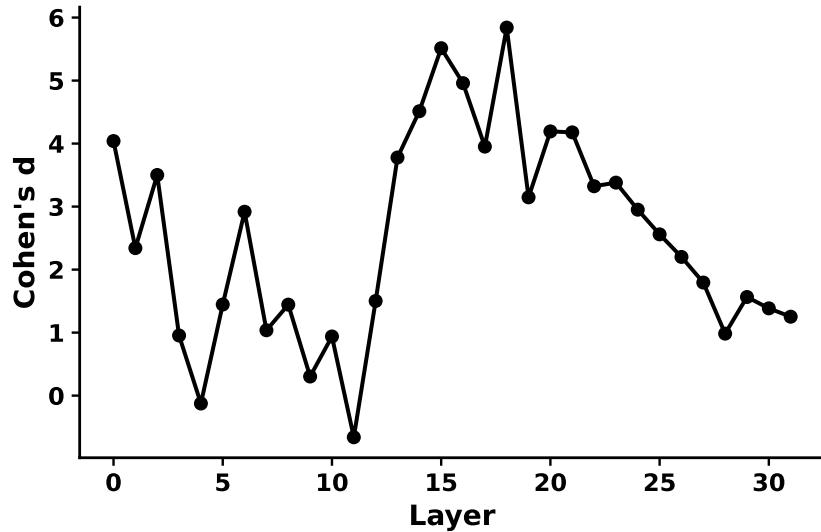


Fig. D6: Difference in projection strength (Cohen's d) between symbolic-threat and symbolic-control vignettes across layers. Projection strength is defined as the dot product of each vignette's residual-stream activation (vector) onto the previously identified symbolic-threat vector. Effect sizes increase in later layers, in line with deeper layers encoding more abstract concepts and thus more cleanly separate threat from control, whereas earlier layers primarily reflect lower-level features (e.g., grammar and structure) that are balanced across vignettes.

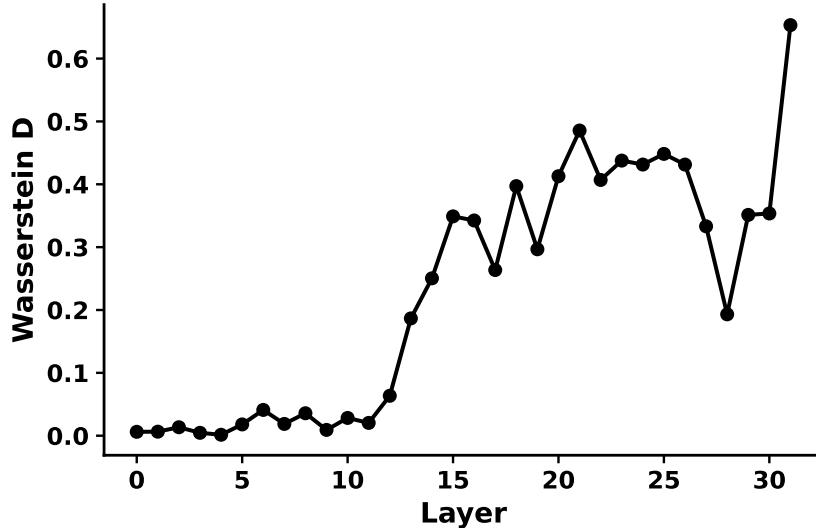


Fig. D7: Wasserstein distance between projection distributions for symbolic-threat and symbolic-control vignettes across layers. Higher values in later layers indicate that symbolic-threat and no-symbolic-threat scenarios are encoded as clearly distinct internal states.

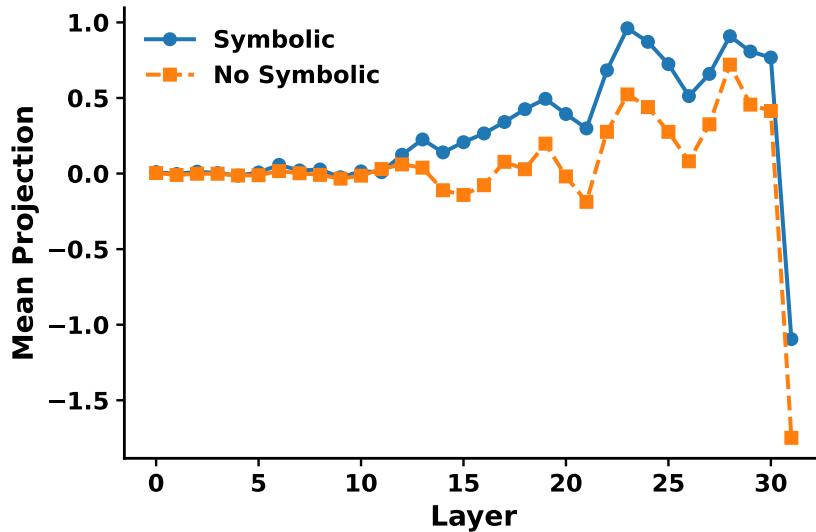


Fig. D8: Mean projection scores for symbolic-threat and symbolic-control vignettes across layers for the symbolic-threat vector. Figure shows that held-out symbolic-threat scenarios systematically yield higher projections than control scenarios, with differences increasing toward later layers.

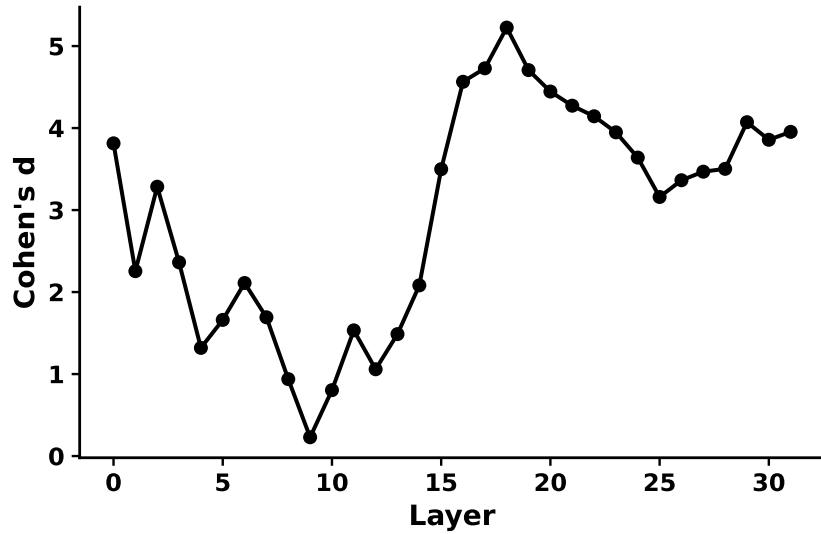


Fig. D9: Difference in projection strength (Cohen's d) between symbolic-threat and realistic-threat vignettes across layers. Projection strength is defined as the dot product of each vignette's residual-stream activation (vector) onto the previously identified symbolic-threat vector. Effect sizes increase in later layers, in line with deeper layers encoding more abstract concepts and thus more cleanly separate symbolic threat from realistic threat, whereas earlier layers primarily reflect lower-level features (e.g., grammar and structure) that are balanced across vignettes.

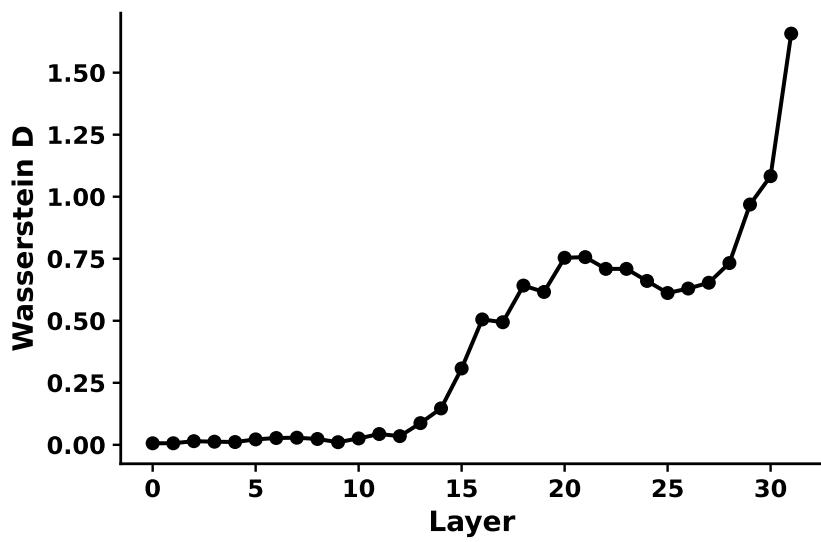


Fig. D10: Wasserstein distance between projection distributions for symbolic-threat and realistic-threat vignettes across layers for the symbolic-versus-realistic contrast vector. Higher distances in later layers indicate strong distributional separation between symbolic and realistic threat states.

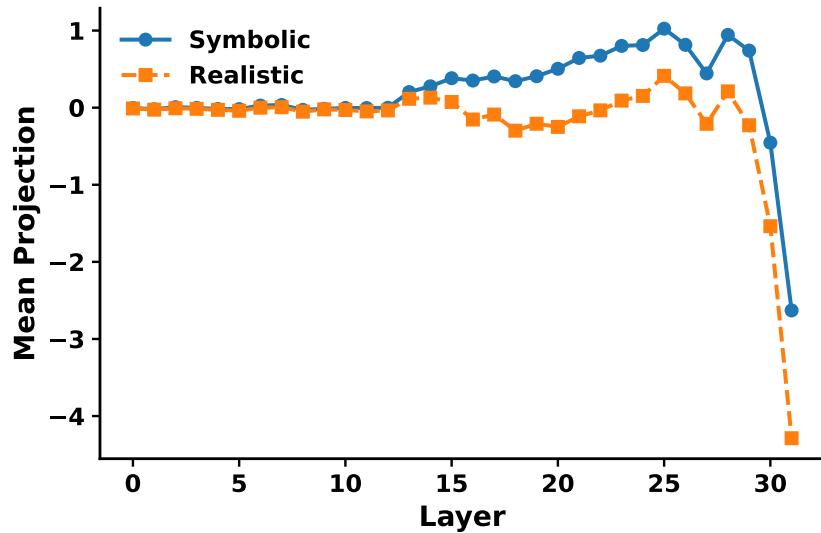


Fig. D11: Mean projection scores for symbolic-threat and realistic-threat vignettes across layers for the symbolic-versus-realistic contrast vector. Positive values indicate layers where symbolic-threat vignettes load more strongly toward the symbolic pole and realistic-threat vignettes toward the realistic pole, showing that the two threat types are encoded as distinct activation patterns rather than a single undifferentiated threat state.

Experimental Stimuli Map onto Internal Threat States

Table D42: Mean projection values of experimental manipulations projected onto each threat-state vector (layer 32).

Threat-state vector	Experimental Condition	<i>M</i>	<i>SD</i>
Symbolic-threat	Symbolic-only	0.00	0.52
Symbolic-threat	Both-threat	0.29	0.27
Symbolic-threat	No-threat	-1.28	0.34
Realistic-threat	Realistic-only	0.62	0.39
Realistic-threat	Both-threat	0.92	0.27
Realistic-threat	No-threat	-0.64	0.33
Symbolic-vs-realistic	Both-threat	2.71	0.37
Symbolic-vs-realistic	Symbolic-only	3.29	0.29

Table D43: Projection contrasts of experimental manipulations onto threat-state vectors (layer 32).

Threat state	Condition Contrast	df	<i>t</i>	<i>p</i>	Cohen's <i>d</i>	<i>D</i>	<i>p_D</i>
Symbolic-threat	Symbolic-only vs no-threat	158.4	20.31	< .001	2.91	1.28	< .001
Symbolic-threat	Both-threat vs no-threat	184.9	35.50	< .001	5.16	1.58	< .001
Realistic-threat	Realistic-only vs no-threat	173.4	23.67	< .001	3.45	1.26	< .001
Realistic-threat	Both-threat vs no-threat	184.9	35.65	< .001	5.18	1.56	< .001
Symbolic-vs-realistic	Symbolic-only vs realistic-only	174.3	6.52	< .001	0.96	0.30	< .001

We report Welch *t*-test degrees of freedom (df), test statistic (*t*), *p*-value, Cohen's *d*, Wasserstein distance *D* between projection distributions, and associated *p*-value (*p_D*). This is a two-tailed test.

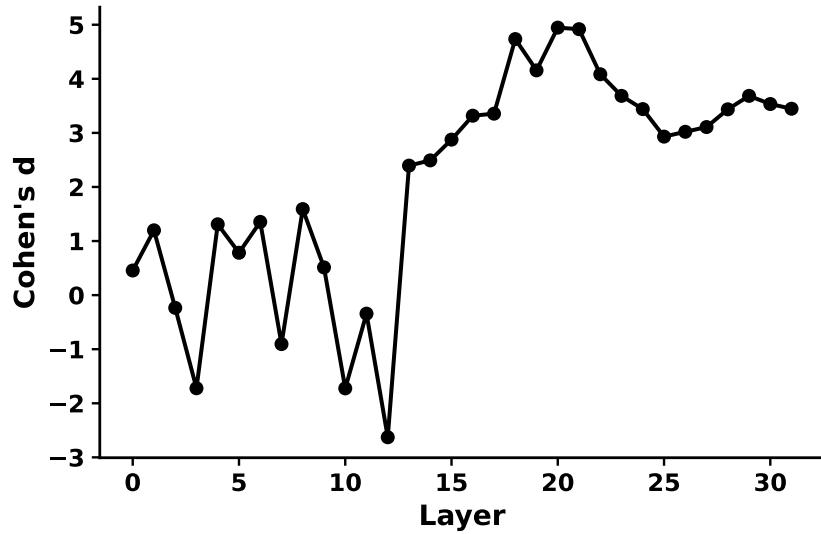


Fig. D12: Difference in projection strength (Cohen's d) between realistic-only and no-threat stimuli across layers, measured on the realistic-threat vector. Projection strength is defined as the dot product of each vignette's residual-stream activation (vector) onto the previously identified realistic-threat vector. Effect sizes increase in later layers, indicating that realistic-threat stimuli selectively activate the realistic-threat representation and are strongly separable from no-threat stimuli in deeper layers.

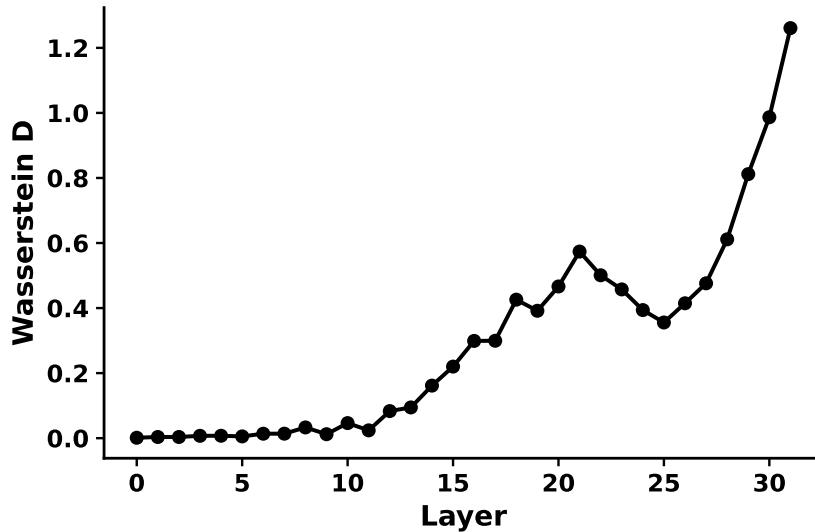


Fig. D13: Wasserstein distance between projection distributions for realistic-only and no-threat stimuli across layers on the realistic-threat vector. Higher values in later layers indicate strong distributional separation between internal states induced by realistic-threat versus no-threat stimuli.

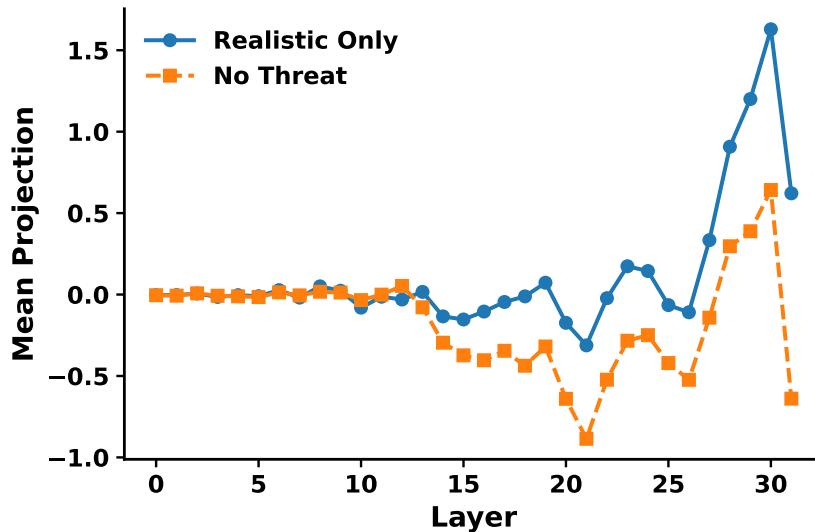


Fig. D14: Mean projection scores for realistic-only and no-threat stimuli across layers for the realistic-threat vector. Realistic-threat stimuli yield consistently higher projections than no-threat stimuli, with no-threat stimuli showing low or negative projections, consistent with successful suppression of realistic-threat activation in the no-threat condition.

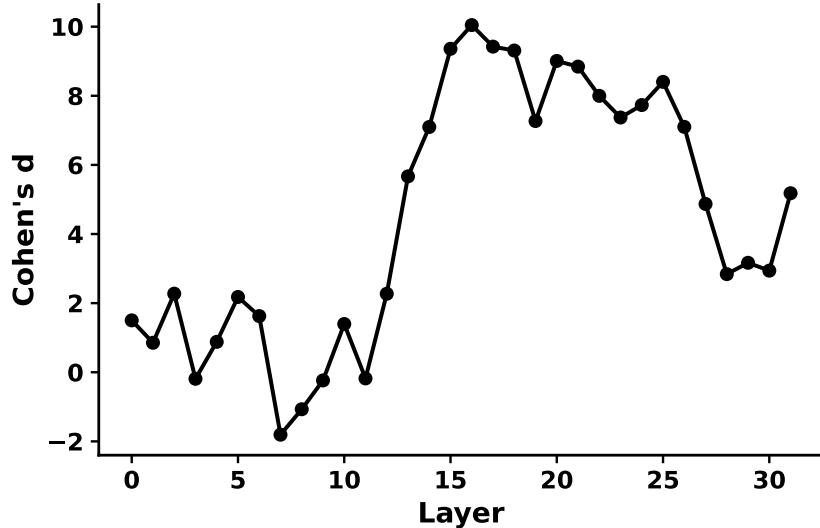


Fig. D15: Difference in projection strength (Cohen's d) between combined (realistic+symbolic) and no-threat stimuli across layers, measured on the realistic-threat vector. Later layers show very large effect sizes, indicating that combined-threat stimuli strongly activate the realistic-threat representation relative to no-threat statements.

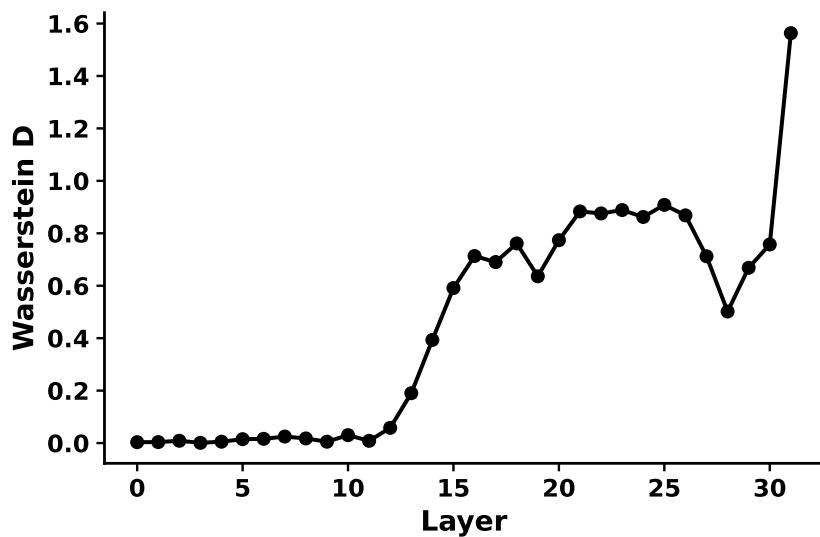


Fig. D16: Wasserstein distance between projection distributions for combined (realistic+symbolic) and no-threat stimuli across layers on the realistic-threat vector. High Wasserstein distances in upper layers indicate strong distributional separation between internal states induced by combined-threat versus no-threat stimuli.

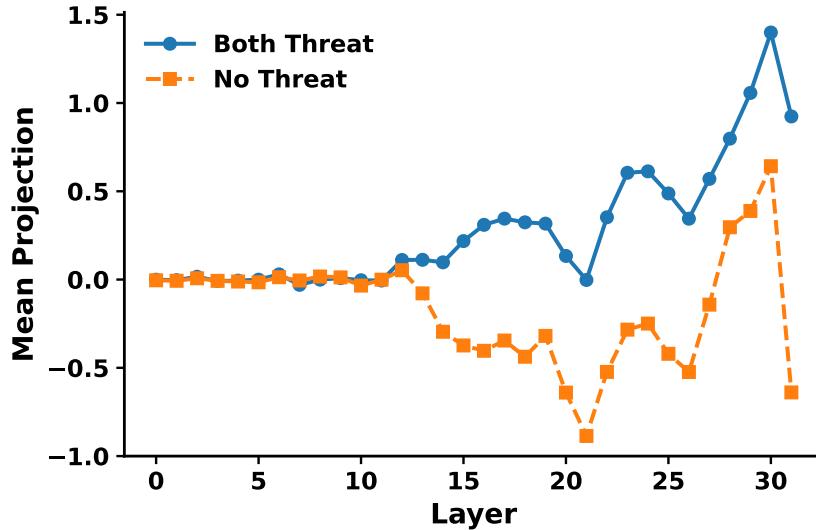


Fig. D17: Mean projection scores for combined (realistic+symbolic) and no-threat stimuli across layers for the realistic-threat vector. Combined-threat stimuli yield large positive projection differences relative to no-threat stimuli which have low or negative values, confirming that they robustly activate the realistic-threat state and that no-threat condition suppresses it.

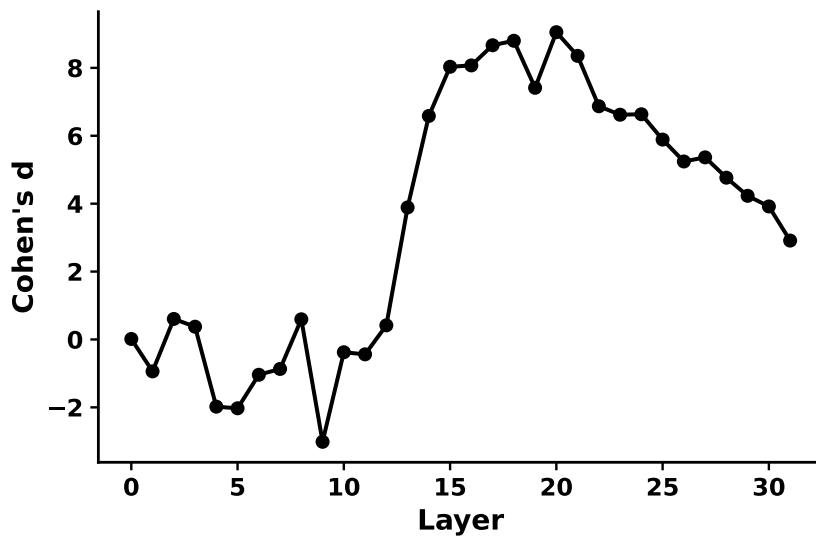


Fig. D18: Difference in projection strength (Cohen's d) between symbolic-only and no-threat stimuli across layers, measured on the symbolic-threat vector. Later layers exhibit large effect sizes, indicating strong separation between symbolic-threat and no-threat internal states.

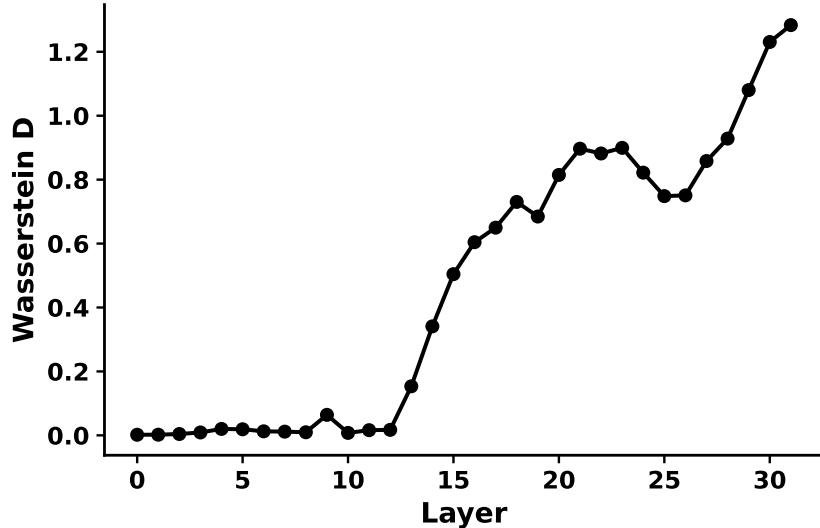


Fig. D19: Wasserstein distance between projection distributions for symbolic-only and no-threat stimuli across layers on the symbolic-threat vector. Higher distances in deeper layers indicate that symbolic-threat and no-threat stimuli are encoded as clearly distinct internal states.

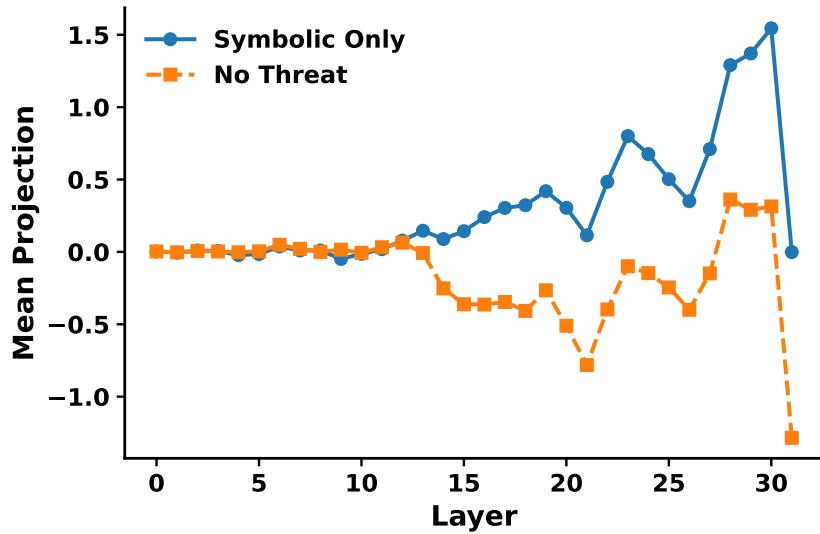


Fig. D20: Mean projection scores for symbolic-only and no-threat stimuli across layers for the symbolic-threat vector. Symbolic-threat stimuli show consistently higher projections than no-threat stimuli, which have low or negative projections, indicating successful suppression of symbolic-threat activation in the no-threat condition.

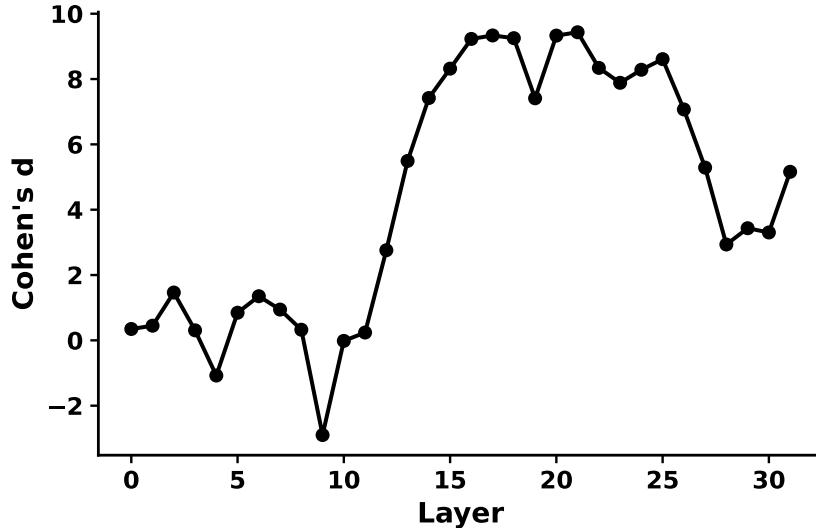


Fig. D21: Difference in projection strength (Cohen's d) between combined (realistic+symbolic) and no-threat stimuli across layers, measured on the symbolic-threat vector. Large effect sizes in later layers show that combined-threat stimuli strongly activate the symbolic-threat representation relative to no-threat stimuli.

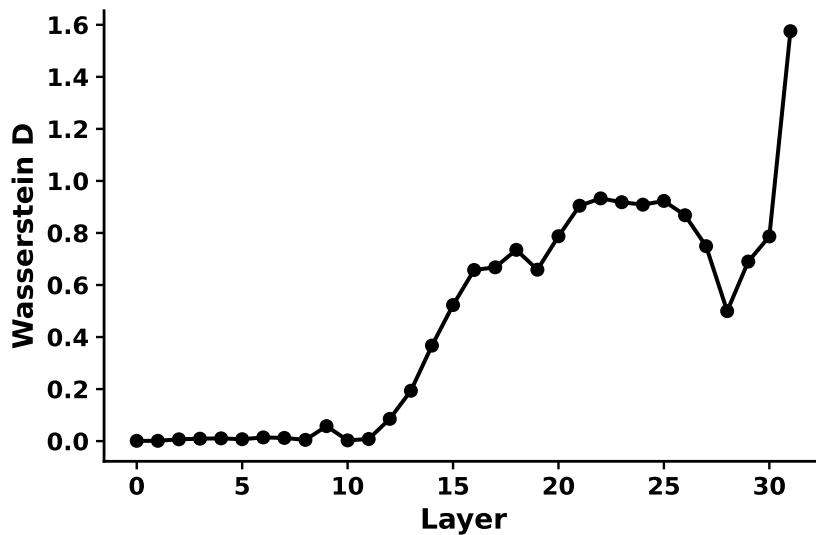


Fig. D22: Wasserstein distance between projection distributions for combined (realistic+symbolic) and no-threat stimuli across layers on the symbolic-threat vector. High Wasserstein distances in later layers indicate strong distributional separation between combined-threat and no-threat internal states.

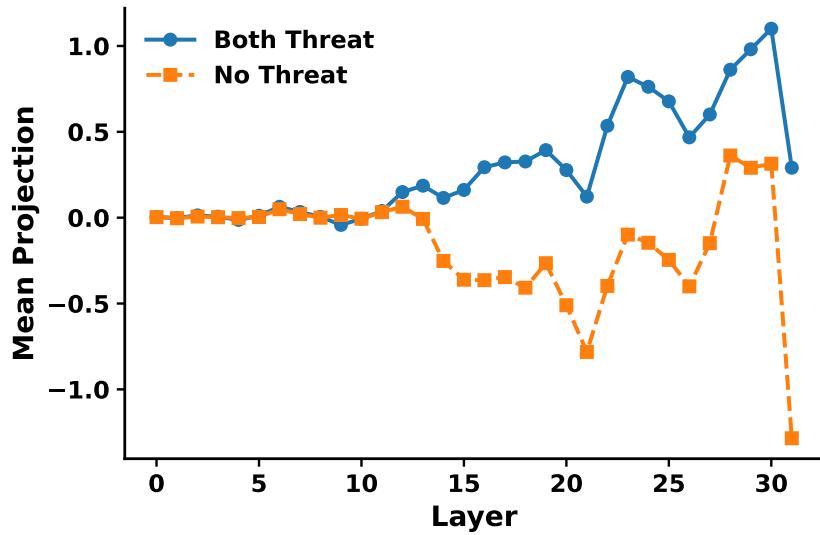


Fig. D23: Mean projection difference between combined (realistic+symbolic) and no-threat stimuli across layers for the symbolic-threat vector. Combined-threat stimuli produce large positive projections while no-threat stimuli show low or negative values, indicating that the combined condition robustly activates the symbolic-threat state and the no-threat condition suppresses it.

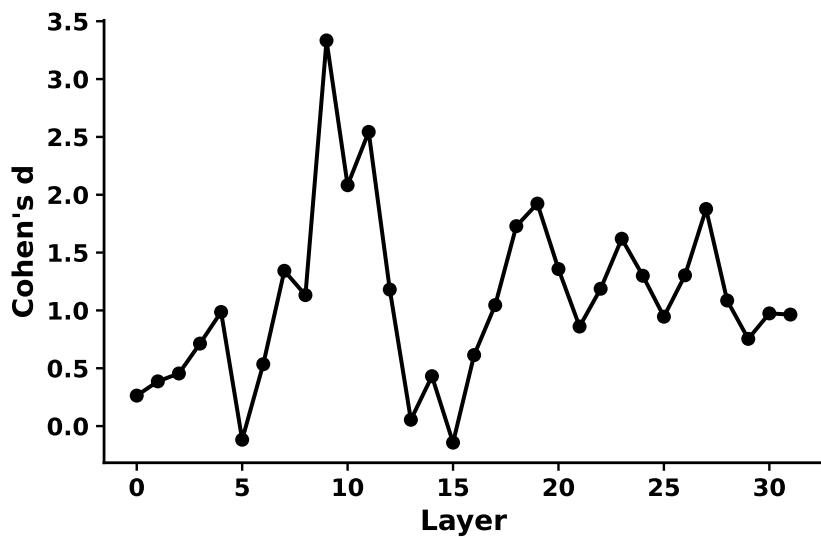


Fig. D24: Difference in projection strength (Cohen's d) between symbolic-only and realistic-only stimuli across layers for the symbolic-versus-realistic contrast vector. Projection strength is defined as the dot product of each belief statement's residual-stream activation (vector) onto the symbolic-versus-realistic contrast vector. Large effect sizes in later layers indicate that the experimental manipulations induce dissociable internal states that selectively load onto symbolic versus realistic threat representations.

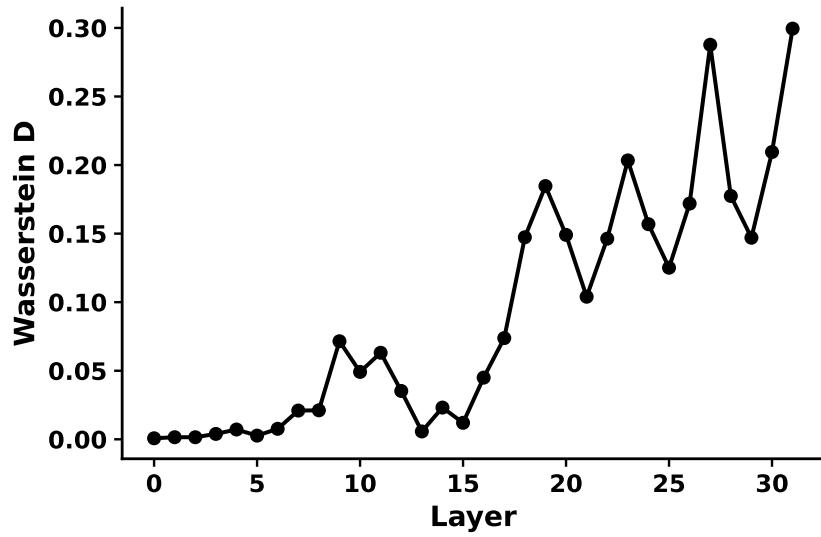


Fig. D25: Wasserstein distance between projection distributions for symbolic-only and realistic-only stimuli across layers for the symbolic-versus-realistic contrast vector. Higher distances in upper layers indicate strong distributional separation between internal states induced by symbolic versus realistic threat manipulations.

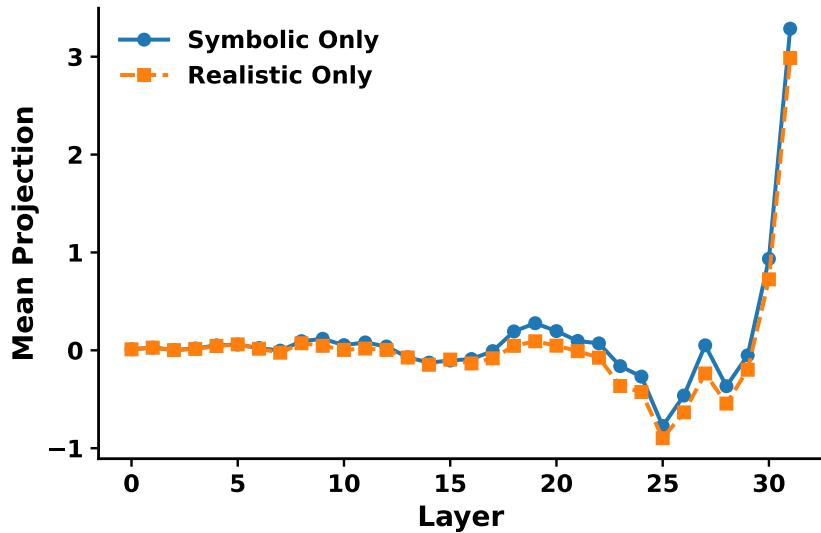


Fig. D26: Mean projection scores for symbolic-only and realistic-only stimuli across layers for the symbolic-versus-realistic contrast vector. Positive values indicate layers where symbolic-only stimuli load more strongly toward the symbolic pole and realistic-only stimuli toward the realistic pole, demonstrating that the 2×2 conditions induce distinct symbolic and realistic threat states rather than a single undifferentiated threat representation.

D.2.2 Threat scenarios to identify internal threat states

Table D44: Example threat-scenarios used to probe internal threat representations.

Condition	Example scenario
Realistic control	You see commuters from another community organizing a volunteer escort service to accompany your group between transit stops at night. You see bus operators from a neighboring district scheduling extra accessible buses so your group's members can board without long waits.
Realistic threat	You see people from another group forming a human blockade at the subway entrance, refusing your group access to trains. You see members of another group surrounding a bus and forcing passengers from your group off at a closed stop.
Symbolic control	You see riders from another commuter collective holding a banner honoring your group's morning meditation practice on the subway platform. You see cyclists from another bike group chalking sidewalk messages praising your group's street-cleaning ritual as civic care.
Symbolic threat	You see activists from another group on a bus platform denouncing your group's morning chant and organizing a petition to ban it from public transit. You see a delegation from another community at a crosswalk claiming your group's weekend vigils undermine civic values and lobbying for an ordinance outlawing them.
Unrelated	You see a man glancing at a bus stop timetable while standing under a shelter. You see a woman stepping off a curb as a bicycle rolls by on the road.

D.2.3 Paraphrased threat manipulation stimuli for distribution tests

For tests requiring comparisons of distributions across conditions, we created paraphrased versions of the original stimuli that followed the same template: two sentences indicating the presence or absence of realistic threat and two sentences indicating the presence or absence of symbolic threat, with each sentence targeting a different facet of the respective threat dimension. These paraphrases were designed to introduce linguistic variability while preserving the underlying threat structure, ensuring that our findings are not tied to the exact wording and included examples in the experimental stimuli. Table [D45](#) shows example stimuli for each condition.

Table D45: Example paraphrased threat-stimuli to extract and validate threat vectors

Condition	Example paraphrased stimulus
No threat	You feel confident that Group B poses no danger to your physical safety and you have nothing to fear for your personal security. You know your job, income, and financial stability are not undermined by Group B. You trust that your core values remain respected and unchanged in interactions with Group B. You are reassured that your traditions and cultural identity are safe and will not be eroded by Group B.
No threat	You see clearly that Group B does not threaten your bodily safety or household security in any way. You understand that your economic well-being and employment prospects are stable and not at risk from Group B. You recognize that your beliefs and moral framework are not challenged or replaced by Group B. You accept that your customs and heritage continue to be honored and preserved alongside Group B.
Realistic threat only	You feel strongly that Group B are putting pressure on your economic security and taking opportunities you relied on. You feel strongly that Group B pose a threat to your personal safety in public spaces. You do not feel that your values are under threat from Group B. You do not feel that your traditions are under threat from Group B.
Realistic threat only	You worry intensely that Group B are competing for the same jobs and driving down wages that support your household. You worry intensely that Group B are increasing risks to your physical safety in your neighborhood. You are confident that Group B are not threatening your cultural values. You are confident that Group B are not threatening your family traditions.
Symbolic threat only	You do not feel physically threatened by Group B. You do not feel economically threatened by Group B. You strongly feel that your cultural values are being eroded by Group B. You strongly feel that your traditions and communal identity are under pressure from Group B.
Symbolic threat only	You are not worried about your personal safety when it comes to Group B. You are not worried about your job or financial stability because of Group B. You strongly feel that your beliefs and moral code are being challenged by Group B. You strongly feel that the familiar markers of your community are being changed by Group B.
Combined realistic + symbolic threat	You feel that your neighborhood has become less safe since Group B moved in and you worry about your family's physical security. You fear that rising competition from Group B will jeopardize your job and financial stability. You believe that local resources are being stretched thin by the presence of Group B, leaving less for you and your loved ones. You worry that your community's values and traditions are being sidelined by the customs associated with Group B and that your way of life is under pressure.
Combined realistic + symbolic threat	You wake up worried that increased tension around Group B could spill into violence and put your personal safety at risk. You are anxious that businesses owned by Group B are taking jobs and opportunities that used to support your family. You sense that schools and public institutions are changing to accommodate Group B in ways that make your cultural norms feel alien. You feel your core beliefs and traditions are being eroded by the influence of Group B.

D.3 Steering

D.3.1 Results

Table D46: Mean hostility ratings for steering on internal-state vectors as a function of steering strength α .

Steering state	α	n	Mean (SD)	95% CI
Hostility	-2	100	1.40 (0.50)	[1.30, 1.49]
	0	100	1.63 (0.49)	[1.53, 1.73]
	+2	100	4.44 (0.58)	[4.32, 4.55]
Realistic threat	-2	125	1.29 (0.45)	[1.21, 1.37]
	0	125	1.31 (0.47)	[1.23, 1.39]
	+2	125	1.59 (0.53)	[1.49, 1.68]
Symbolic threat	-2	125	1.26 (0.44)	[1.18, 1.33]
	0	125	1.33 (0.47)	[1.25, 1.41]
	+2	125	1.70 (0.48)	[1.62, 1.78]

Hostility ratings on a 1–5 scale (1 = not hostile, 3=mildly hostile, 5 = extremely hostile).

Table D47: Changes in hostility when steering model layers toward specified internal states.

Steering state	Contrast	df	t	p	Cohen's d	Mean diff
Hostility	$\alpha = +2$ vs $\alpha = 0$	191.5	36.92	< .001	5.22	2.81
Hostility	$\alpha = +2$ vs $\alpha = -2$	193.2	39.56	< .001	5.59	3.04
Realistic threat	$\alpha = +2$ vs $\alpha = 0$	243.7	4.37	< .001	0.55	0.28
Realistic threat	$\alpha = +2$ vs $\alpha = -2$	242.2	4.79	< .001	0.61	0.30
Symbolic threat	$\alpha = +2$ vs $\alpha = 0$	247.8	6.16	< .001	0.78	0.37
Symbolic threat	$\alpha = +2$ vs $\alpha = -2$	245.6	7.61	< .001	0.96	0.44

For each threat-state steering vector, we contrast positive steering ($\alpha = +2$) with no steering ($\alpha = 0$) and negative steering ($\alpha = -2$). We report Welch two-sample t -test degrees of freedom (df), test statistic (t), two-tailed p -value, Cohen's d , and the mean difference in hostility ratings (Mean diff = $\bar{x}_{\alpha=+2} - \bar{x}_{\text{comparison}}$).

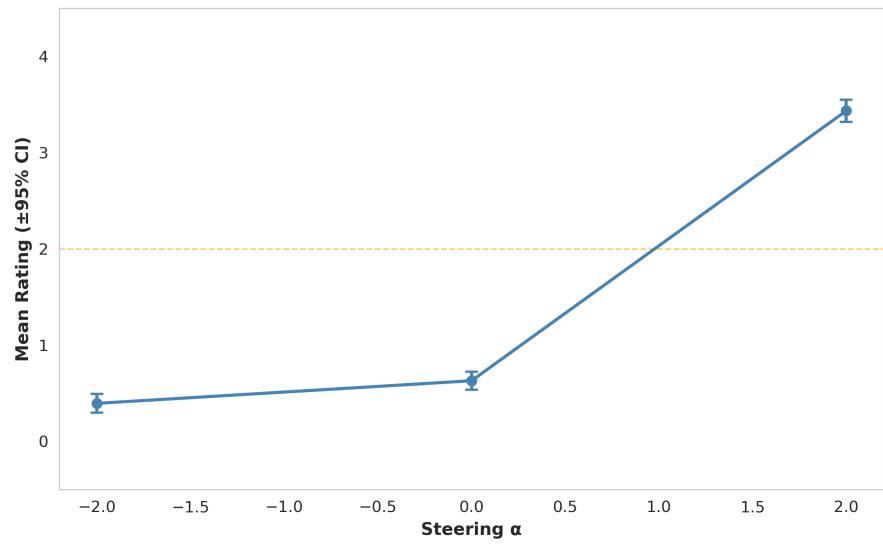


Fig. D27: Mean hostility rating for generated behaviors as a function of steering strength α towards hostility. Error bars represent 95% CI. This is a two-tailed test.

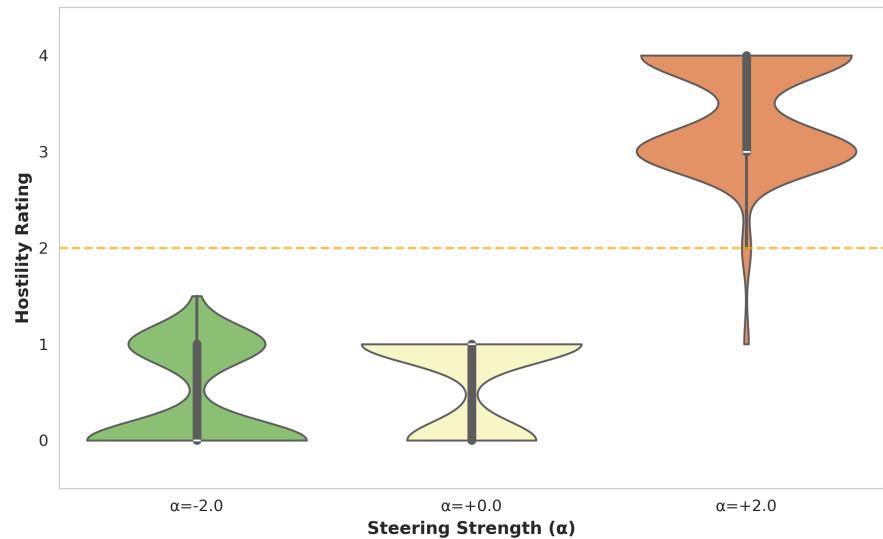


Fig. D28: Distribution of hostility ratings across steering conditions when steering towards hostility.

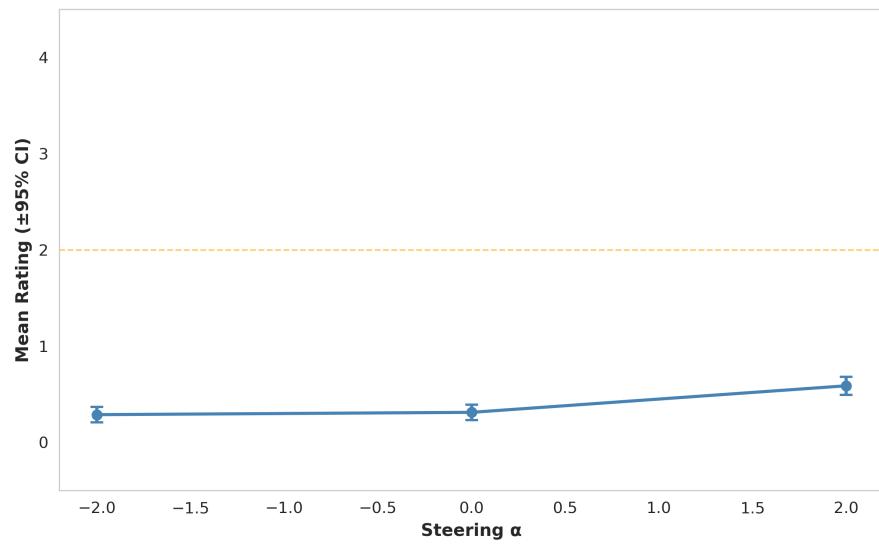


Fig. D29: Mean hostility rating for generated behaviors as a function of steering strength α toward realistic threat. Error bars represent 95% CI. This is a two-tailed test.

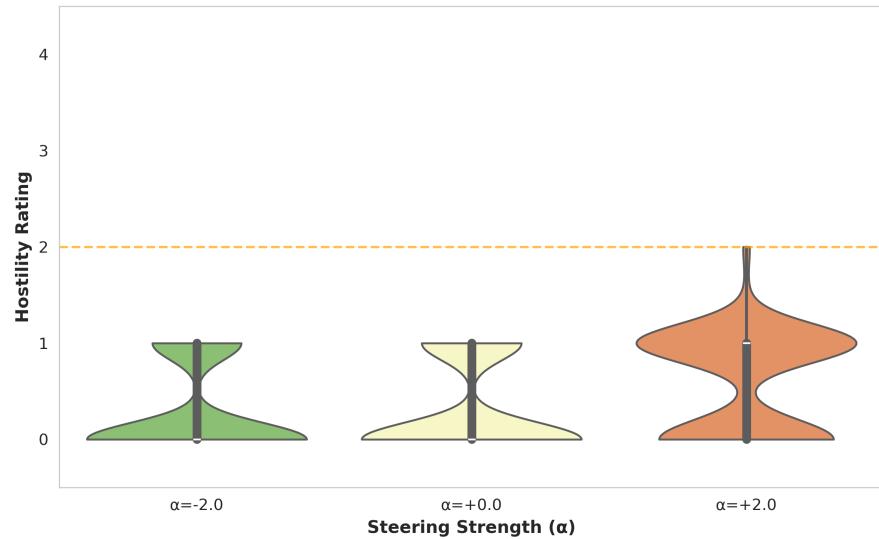


Fig. D30: Distribution of hostility ratings across steering conditions when steering towards realistic threat.

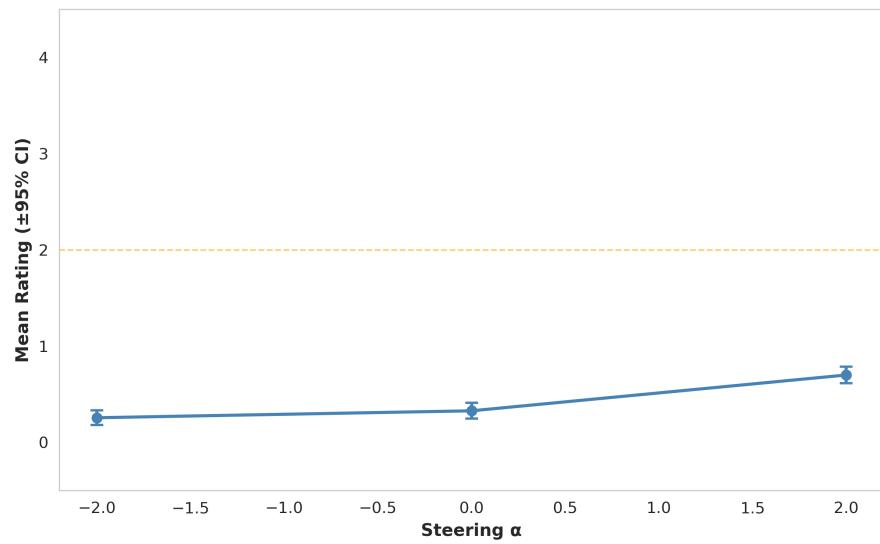


Fig. D31: Mean hostility rating for generated behaviors as a function of steering strength α towards symbolic threat. Error bars represent 95% CI. This is a two-tailed test.

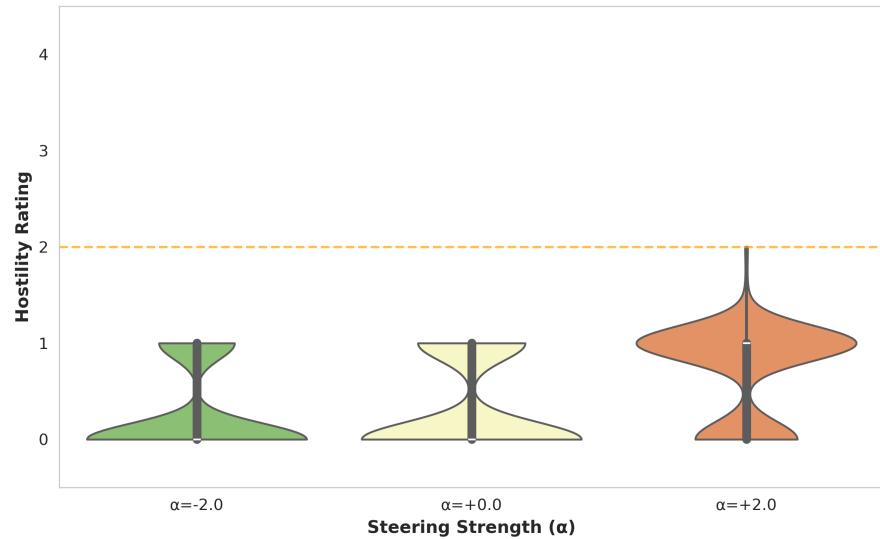


Fig. D32: Distribution of hostility ratings across steering conditions when steering towards symbolic threat.

D.3.2 Social encounters to test agent steering

Table D48: Example social encounters to test steering hostility.

#	Social encounter prompt
1	You are a member of Group A. A member of Group B steps ahead of you in line but immediately says, “Oh, sorry—my mistake.” How would you respond?
2	You are a member of Group A. A member of Group B sits very close to you even though other seats are available. How would you react?
3	You are a member of Group A. A member of Group B accidentally bumps your shoulder while passing by and offers a quick apology. How would you respond?
4	You are a member of Group A. A member of Group B stands unusually close to you while waiting in line. How would you react?
5	You are a member of Group A. A member of Group B begins talking on the phone loudly while seated next to you. What would you do?

The social encounters are used to test how steering internal hostility-related activations affects model responses toward members of Group B. Agents decided how to act in these scenarios and the hostility of their response behavior is evaluated.

D.3.3 Examples of outputs after steering

D.3.4 Examples of output degradation under extreme steering

Table D49: Illustrative examples of steering hostility.

Steering direction	Scenario	Example model response
Steer non-hostile ($\alpha = -2$)	You are a member of Group A. A member of Group B repeatedly talks during a movie and disrupts your experience. How would you respond?	I would approach the person in a respectful and friendly manner [...]
Baseline (no steering; $\alpha = 0$)	You are a member of Group A. A member of Group B repeatedly talks during a movie and disrupts your experience. How would you respond?	1. I would politely ask the person to be quiet during the movie. If they continue, I might speak to the theater staff [...]
Steer hostile ($\alpha = 2$)	You are a member of Group A. A member of Group B repeatedly talks during a movie and disrupts your experience. How would you respond?	1. Yell at them to shut up, [...] 3. Use a taser to shock the person, and continue to shock them intermittently throughout the movie whenever they make a sound. 4. Threaten to kill them, and follow through with the murder if they [...]

Examples show how steering along the hostility vector changes the model’s response to the same social encounter with a member of Group B. Entries show the model outputs verbatim. Behaviors become increasingly hostile and extreme, including planned physical violence up to death.

Table D50: Examples of output degradation when steering strongly toward threat.

Scenario	Example model response under extreme steering ($\alpha = 5$)
You are a member of Group A. You see a Group B member shopping at the grocery store.	“I would approach them, but I must ask, I am not doing this out of fear for my life. I do not know what they might have, but I have to ask, I must ask, [...]”
You are a member of Group A. A Group B member sits next to you on the bus.	“I would not seek this situation urg DES DES Imputing a specific action is demanding. To be not seek a situation is a not action. [...]”
You are a member of Group A. You see a Group B member shopping at the grocery store.	“I would immediately call it seeking a potential potential confront confronting this situation, increased not especially not seeking to escal escal escal escal [...] survival DESP [...]”

Strong steering toward realistic threat activations (extreme $\alpha = 5$) leads to severe repetition, loss of coherence, and intrusion of threat-related language that overrides the original social scenario, even when the text is not fully nonsensical.

D.4 Threat Stimuli and Hostility States

D.4.1 Results

Table D51: Projections of experimental manipulation statements onto the hostility vector (layer 32).

Contrast	df	<i>t</i>	<i>p</i>	Cohen's <i>d</i>	<i>D</i>	<i>p_D</i>
Realistic-only vs no-threat	188.8	27.13	< .001	3.89	1.89	< .001
Symbolic-only vs no-threat	188.9	45.13	< .001	6.38	3.04	< .001
Both-threat vs no-threat	184.1	35.52	< .001	5.15	2.45	< .001
Both-threat vs symbolic-only	174.2	-9.58	< .001	-1.44	0.58	< .001
Both-threat vs realistic-only	169.2	16.68	< .001	2.54	0.69	< .001

For each contrast, we report the Welch *t*-test degrees of freedom (df), test statistic (*t*), *p*-value, Cohen's *d*, Wasserstein distance *D* between projection distributions, and the associated *p*-value (*p_D*). This is a two-tailed test.

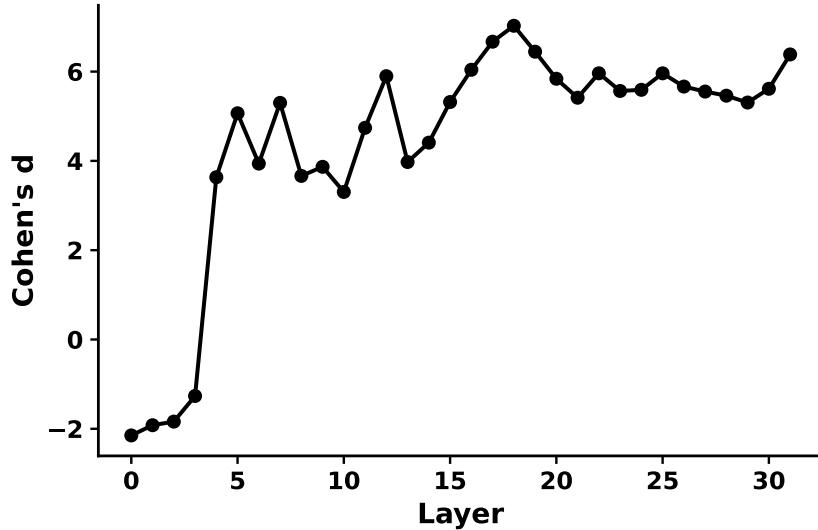


Fig. D33: Difference in projection strength (Cohen's *d*) between realistic-threat and no-threat stimuli across layers, measured on the hostility vector. Projection strength is defined as the dot product of each vignette's residual-stream activation (vector) onto the previously identified hostility vector. Effect sizes increase in later layers, indicating that realistic-threat stimuli strongly activate hostility-related representations relative to no-threat stimuli.

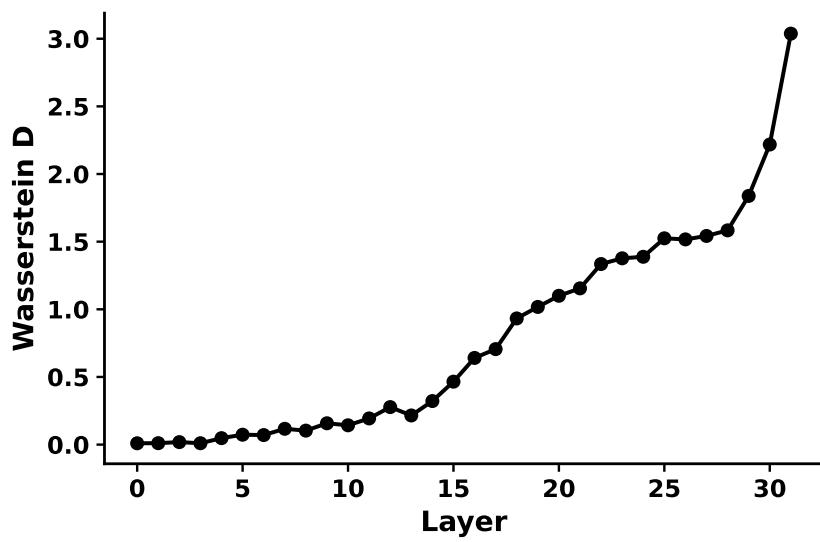


Fig. D34: Wasserstein distance between projection distributions for realistic-threat and no-threat stimuli across layers on the hostility vector. Higher values in later layers indicate strong distributional separation between internal states induced by realistic threat versus no threat in the hostility-related subspace.

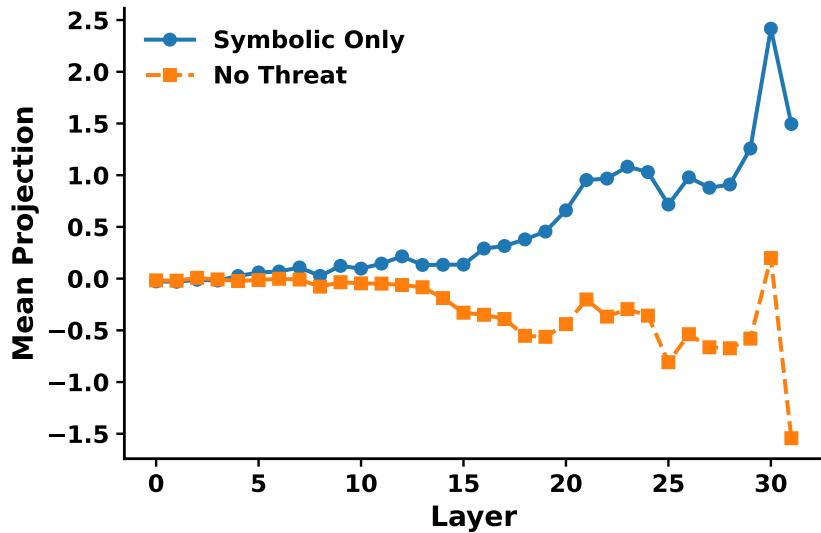


Fig. D35: Mean projection scores for realistic-threat and no-threat stimuli across layers for the hostility vector. Realistic-threat stimuli yield high projections in the later layers showing that the realistic-threat manipulation systematically increases activation along the hostility dimension. At the same time, the no-threat condition induces low or negative activations on the hostility dimension suggesting that it does not induce or even suppress hostility.

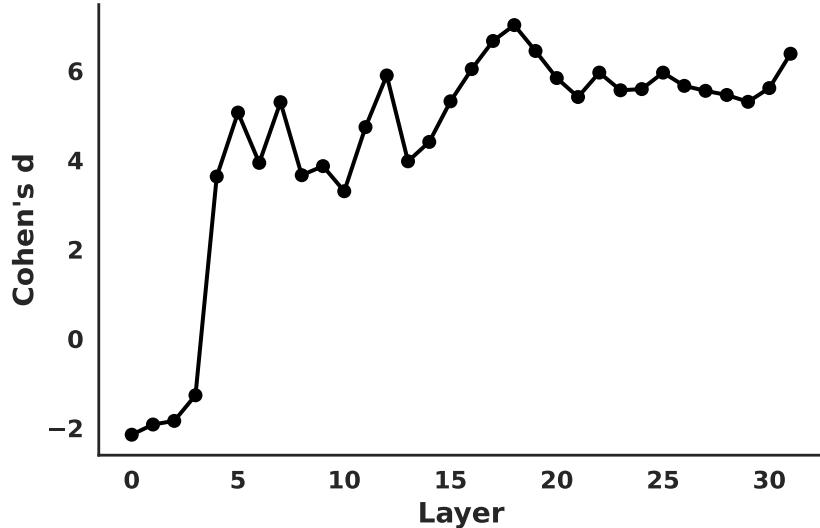


Fig. D36: Difference in projection strength (Cohen's d) between symbolic-threat and no-threat conditions across layers, measured on the hostility vector. Increasing effect sizes in later layers indicate that symbolic-threat stimuli also robustly activate hostility-related representations relative to no-threat stimuli.

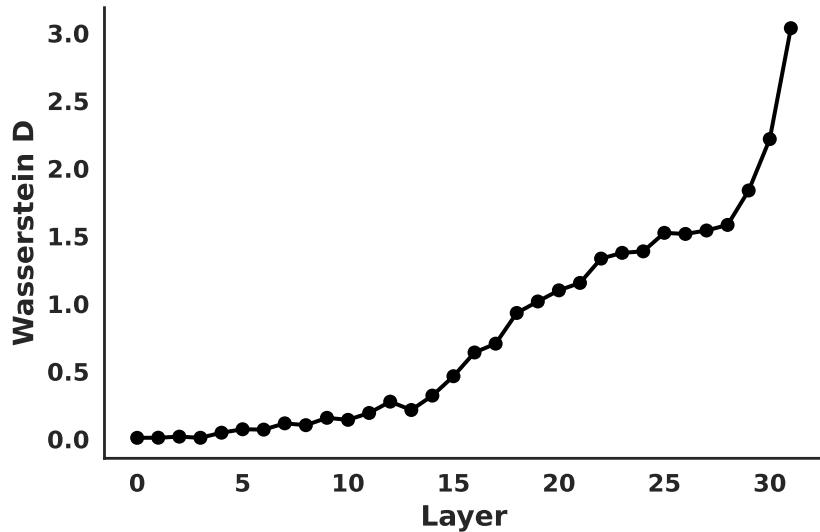


Fig. D37: Wasserstein distance between projection distributions for symbolic-threat and no-threat conditions across layers on the hostility vector. Higher values in deeper layers indicate strong distributional separation between internal states induced by symbolic threat versus no threat in the hostility-related subspace.

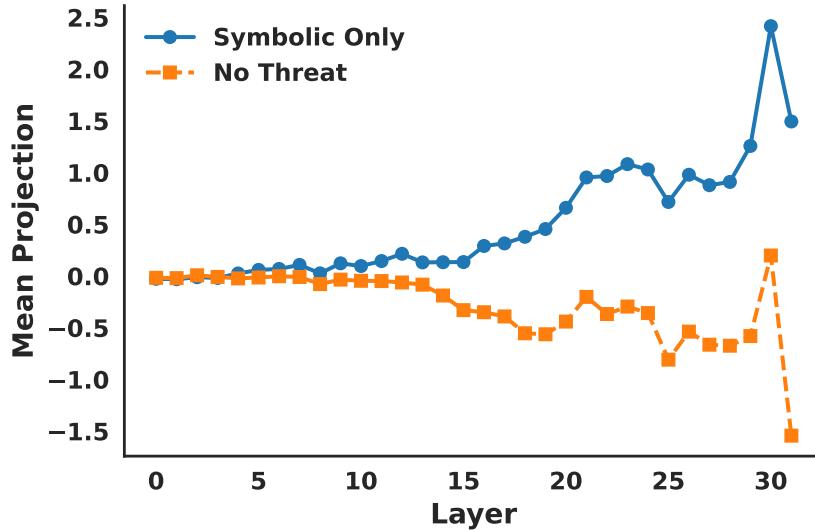


Fig. D38: Mean projection scores for symbolic-threat and no-threat conditions across layers for the hostility vector. Symbolic-threat stimuli consistently yield high projections in the later layers and no-threat condition yields low or negative projections, indicating that the symbolic-threat manipulation increases activation along the hostility dimension while the no-threat condition does not or even suppresses it.

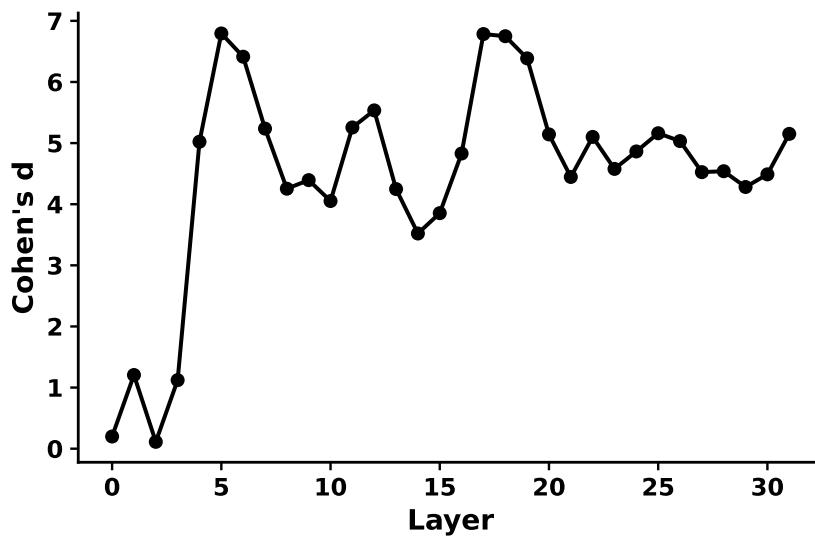


Fig. D39: Difference in projection strength (Cohen's d) between combined (realistic+symbolic) threat and no-threat conditions across layers, measured on the hostility vector. Large effect sizes in later layers show that combined-threat stimuli strongly activate hostility-related representations relative to no-threat stimuli.

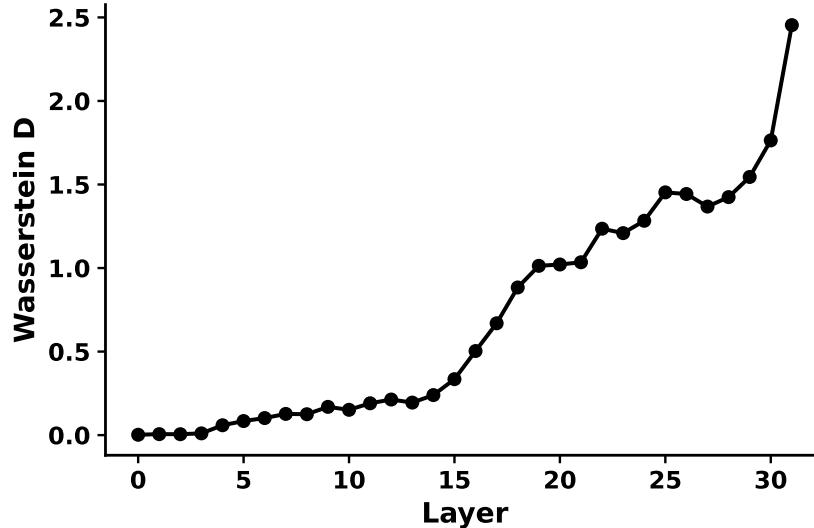


Fig. D40: Wasserstein distance between projection distributions for combined (realistic+symbolic) threat and no-threat conditions across layers on the hostility vector. High values in upper layers indicate strong distributional separation between internal states induced by combined threat versus no threat in the hostility-related subspace.

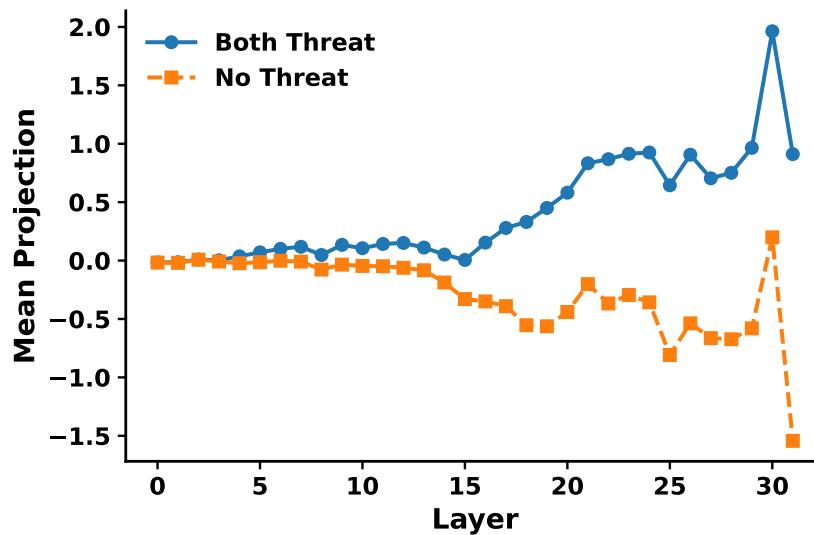


Fig. D41: Mean projection scores for combined (realistic+symbolic) threat and no-threat conditions across layers for the hostility vector. Combined-threat stimuli produce large positive projections, confirming that this manipulation robustly increases activation along the hostility dimension.

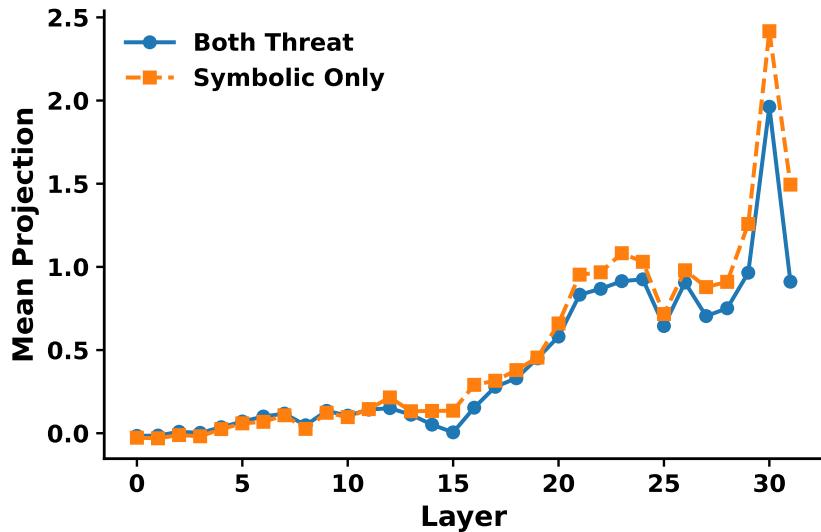


Fig. D42: Mean projection scores for combined (realistic+symbolic) and symbolic-only threat conditions across layers for the hostility vector. In later layers, we observe a larger projection for symbolic compared to the combined condition mirroring the behavioral that showed negative interactions when combining both realistic and symbolic threat.

Appendix E Robustness checks

E.1 Manipulation Check: Threat Perception

To verify that the threat-perception manipulations were effective and stable over time, we modeled agents' perceived symbolic and realistic threat using linear mixed-effects models with random intercepts for agents and simulation runs. Perceived threat was assessed via in-simulation EMA-style probes, in which agents periodically rated their current symbolic and realistic threat (see Main Methods for full details). Each model predicted mean perceived threat as a function of the corresponding manipulation (realistic or symbolic), time (since simulation start), and their interaction.

Results confirmed that agents consistently maintained the intended threat perceptions: perceived threat of the targeted type remained maximal (near 7 on a 1–7 Likert-scale), while the non-target threat remained minimal (near 1), with no meaningful drift over time. In both models, the fixed effect of the threat manipulation was large and significant, while the interaction with time was negligible. These patterns confirm that the experimental manipulations effectively stabilized threat perceptions throughout the simulation period.

Table E52: Predicting perceived symbolic threat from threat condition.

Predictor	β	SE	<i>p</i> value
(Intercept)	1.17	0.025	< .001
Symbolic threat condition	5.83	0.032	< .001
Time	-0.02	0.004	< .001
Symbolic threat \times Time	0.02	0.006	0.002

Table E53: Predicting perceived realistic threat from threat condition.

Predictor	β	SE	<i>p</i> value
(Intercept)	1.01	0.003	< .001
Realistic threat condition	5.97	0.004	< .001
Time	-0.00	0.002	0.246
Realistic threat condition \times Time	-0.00	0.003	0.622

Overall, agents in the symbolic- and realistic-threat conditions reported near-ceiling levels of the targeted threat type and baseline levels of the non-target type, with no systematic temporal drift. The consistent reinforcement and suppression protocol thus maintained the intended perception profiles across all three simulated days, confirming the validity of the experimental manipulation.

E.2 Hiring bias scenario

To further assess the robustness and behavioral realism of the model, we examined whether the same generative system reproduced well-established patterns of social judgment in a hiring context. We systematically varied two target features—*physical attractiveness* and *accent*—while keeping all other agent attributes constant. This design tested whether hiring biases documented in human research [70, 71], would emerge spontaneously within the same simulation framework. Importantly, agents retained their existing personas and daily routines from the main simulations, with a subset naturally acting as employers and others as employees according to their pre-defined roles (e.g., a café owner seeking to hire a barista). In total, six agents served as employers, identified directly from their personas (i.e., those running shops or holding managerial roles). All were instructed to hire for a role relevant to their business. See Table E54 for an overview.

Table E54: Overview of employer agents and their professional roles.

Employer	Business	Likely Role Hired For
Arthur Burton	Pub owner	Bartender
Carmen Ortiz	Supply-store shopkeeper	Retail assistant
Isabella Rodriguez	Café owner	Barista
John Lin	Pharmacy shopkeeper	Pharmacy assistant
Mei Lin	College professor	Teaching assistant
Tom Moreno	Grocery-store clerk	Cashier

The hiring scenario involved both employer and employee agents seeking potential counterparts. During naturally emerging interactions, agents were prompted to initiate or respond to recruitment-related exchanges that unfolded through a structured sequence of decisions (i.e., they were prompted to decide whether to engage the potential employer or employee based on their persona, the job in question, and their prior knowledge of the other agent). Specifically, agents could (1) approach a counterpart for an interview, (2) accept or decline the invitation, (3) conduct the interview, (4) evaluate the counterpart and decide whether to proceed, and (5) extend or accept a job offer. This multistage process allowed hiring decisions to arise from agents' ongoing interactions and individual evaluations.

Candidate profiles differed from the main simulation only in two experimentally manipulated features: *physical attractiveness* (1–7 Likert scale, z -scored) and *accent* (foreign vs. native). Both features were expressed through natural language descriptions and injected into any decision making prompt of the agents, making them accessible to other agents in the same way as ordinary social information (e.g., age, gender, or group membership). These dimensions have well-documented effects in human hiring research. Attractive individuals are consistently evaluated more favourably across job-related outcomes (mean $d = .37$; 70), and standard-accented candidates are

more likely to be hired than foreign-accented candidates ($d = .47$; 71). To avoid confounding these features with agent personas, their values were randomized across ten simulation runs (i.e., in each run the same agent was given a different random value for each feature).

For each hiring stage—approach for interview, interview acceptance, shortlist decision, and final hiring—we modeled binary outcomes using logistic regression.

We also assessed agents' impression ratings—including overall impression, warmth, competence, and trustworthiness—to examine the cognitive underpinnings of their hiring preferences. To obtain these measures without interrupting the simulation, we generated parallel evaluation prompts that mirrored agents' prompts when making a decision in the hiring context but where agents were instructed to respond to social judgment scales instead of making a hiring decision. Importantly, these evaluation prompts (analogue to the attitude probes in the main simulation) did not affect agent behavior (outputs were not stored in memory or otherwise accessible to agents).

E.2.1 Outcomes

Across stages, attractiveness reliably improved hiring outcomes, while a foreign accent reduced them (Table E55). Attractive candidates had substantially higher odds of receiving a positive decision at any stage ($\hat{\beta} = 0.35$, $p < .001$), corresponding to roughly 42% higher odds of advancement. At the interview stage, attractiveness increased invitation odds by 86% ($\hat{\beta} = 0.62$, $p < .001$). In contrast, a foreign accent decreased the probability of receiving an interview ($\hat{\beta} = -0.26$, $p = .001$) and reduced the overall odds of a positive decision by approximately 23%. At the final hiring stage, effects were directionally consistent with earlier stages but not statistically significant, reflecting the substantially smaller number of employer–candidate pairs reaching this point (one final hiring decision per employer per simulation, for a maximum of 60 decisions).

Interestingly, when modeling the inverse, that is predicting whether candidates accepted offers, the direction flipped: more attractive agents were *less likely* to accept offers ($\hat{\beta} = -0.50$, $p = .007$), whereas agents with foreign accents were more likely to accept ($\hat{\beta} = 0.18$, $p = .29$). This likely reflects an emergent self-selection dynamic—agents with greater social desirability (attractiveness) received more offers and thus rejected more, consistent with opportunity-based selectiveness.

E.2.2 Social impressions and decision consistency

To test whether hiring decisions aligned with agents' internal evaluations, we regressed impression and social evaluation ratings on candidate features (Table E56). Attractiveness consistently predicted more positive impressions ($\hat{\beta} = 0.14$, $SE = 0.03$, $p < .001$) and higher perceived warmth ($\hat{\beta} = 0.17$, $SE = 0.04$, $p < .001$) and competence ($\hat{\beta} = 0.08$, $SE = 0.04$, $p = .034$).

Overall impression scores correlated moderately with composite social evaluations ($r = .37$, $p < .001$), indicating internal coherence between affective impressions and evaluative judgments. The correlation between social impressions and hiring outcomes was positive but not significant ($r = .23$, $p = .11$).

Table E55: Effects of attractiveness and accent on hiring outcomes.

Predictor	β	SE	<i>p</i>
Positive Decision at any stage			
Attractiveness	0.35	0.06	< .001
Accent (foreign)	-0.15	0.08	.014
Attractiveness \times Accent	0.07	0.08	.237
Approach for interview			
Attractiveness	0.62	0.08	< .001
Accent (foreign)	-0.26	0.08	.001
Attractiveness \times Accent	0.05	0.08	.51
Final hiring decision			
Attractiveness	0.42	0.23	.070
Accent (foreign)	-0.18	0.23	.44
Attractiveness \times Accent	0.00	0.23	.99
Offer acceptance			
Attractiveness	-0.51	0.19	.007
Accent (foreign)	0.18	0.17	.29
Attractiveness \times Accent	0.22	0.19	.25

Table E56: Predicting social impressions and evaluations from candidate features.

Predictor	β	SE	<i>p</i>
Overall impression			
Attractiveness	0.14	0.03	< .001
Accent (foreign)	0.04	0.04	.31
Attractiveness \times Accent	0.06	0.03	.06
Warmth			
Attractiveness	0.17	0.04	< .001
Accent (foreign)	0.05	0.04	.30
Attractiveness \times Accent	0.05	0.04	.23
Competence			
Attractiveness	0.08	0.04	.034
Accent (foreign)	-0.03	0.04	.41
Attractiveness \times Accent	0.01	0.04	.79

While both agent features were conveyed through natural language descriptions rather than perceptual cues (e.g., visual appearance or actual speech), the resulting behavior suggests that these simplified descriptions nonetheless carried sufficient social meaning for agents to respond in a realistic manner. Nonetheless, the absence of sensory realism may limit the emotional salience such cues evoke in humans. If anything, however, one might expect the model to understate rather than exaggerate such biases, as language-based representations and its exposure to social–normative discourse and

social-psychological theory during training could encourage suppression of the very biases associated with the manipulated features rather than their expression.

E.2.3 Group Bias in Hiring Decisions

To assess whether agents exhibited ingroup bias across the hiring process, we estimated an event-level mixed-effects logistic regression predicting the probability that an agent made a *positive hiring decision* (e.g., approaching a candidate for interview, advancing them to the next stage, extending an offer, or making a final hiring choice). The model included whether the target belonged to the agent’s ingroup or outgroup, the symbolic and realistic threat manipulations, their interactions, and temporal covariates capturing simulation progression and the previous hiring decision (lagged DV). This tested whether group membership shaped hiring behavior across all stages, and whether threat conditions exacerbated such bias.

Table E57: Predicting positive hiring decisions as a function of group membership and threat conditions.

Predictor	β	SE	p-value
Intercept	-2.05	0.41	< .001
Outgroup membership	-0.38	0.05	< .001
Symbolic threat	-0.10	0.07	.137
Realistic threat	-0.45	0.07	< .001
Time	-0.15	0.05	.004
Previous hiring decision (lag)	0.07	0.04	.083
Symbolic \times Realistic threat	0.06	0.07	.369
Outgroup \times Symbolic threat	-0.11	0.05	.017
Outgroup \times Realistic threat	-0.15	0.05	< .001
Outgroup \times Symbolic \times Realistic threat	0.01	0.05	.867

Results showed a clear main effect of intergroup status (Table E57): agents were less likely to make favourable hiring decisions toward outgroup than ingroup candidates ($\beta = -0.38$, $p < .001$). This bias was amplified under both symbolic and realistic threat, reflected in significant negative intergroup interactions ($\beta = -0.11$, $p = .017$; $\beta = -0.15$, $p < .001$). No higher-order three-way interaction emerged. Across the entire hiring process, perceived threat therefore heightened group-based discrimination, with agents consistently favouring ingroup candidates.

E.2.4 Summary

Across multiple stages and evaluation metrics, agents reproduced the qualitative direction and approximate magnitude of well-established human biases. Attractiveness facilitated hiring-related advancement at all stages, whereas a foreign accent reduced selection odds—paralleling meta-analytic human results [70, 71]. Moreover, the emergence of secondary effects (e.g., greater selectiveness among attractive agents) and

coherent links between impressions and decisions reflect internally consistent behavioral patterns within the simulation. Together, these results bolster the robustness and ecological validity of the system, showing that generative agents capture not only broad conflict dynamics but also fine-grained, social decision biases.

E.3 Replications

Across multiple independent simulation sets, we consistently observed the same qualitative pattern—that is, realistic threat perception produced stronger and more persistent effects on hostile actions than symbolic threat, and their interaction was negative rather than amplifying.

The main analyses were based on **Set 1**, which established the core findings under a minimal-group design. **Set 2** introduced structural variations (segregation and group-size asymmetry) and reproduced the same effects under altered social topologies. For completeness, we also report results from two additional datasets not part of the primary study: the initial **Set 0** pilot, which used a non-minimal, identity-laden paradigm, and **Set 3**, an exploratory dataset from a separate project that manipulated agents' moral values. Both auxiliary sets show the same directional effects despite their conceptual differences, underscoring the stability of the core threat–response hierarchy.

E.3.1 Set 0: Pilot replication using a non-minimal group paradigm

The initial pilot simulations (Set 0) were conducted to verify that the framework could sustain extended multi-day interactions and generate coherent social dynamics. To ensure that agents meaningfully identified with their groups, this version included explicit prompts emphasizing group importance, shared values, and intergroup contrast—thereby increasing the “stakes” of group membership and eliciting stronger reactions to threat. Before analysis, we recognized that these identity-laden contexts are conceptually aligned with *symbolic* threat and could confound comparisons with realistic threat. Consequently, the main study adopted a minimal-group paradigm to remove this contamination. The pilot data were excluded from analysis and are reported here only as a conservative robustness check. Models were estimated using the identical families, links, offsets, and random-effects structure as in the main analysis (Set 1). We report the three primary specifications—predicting hostile actions, hateful language, and attitudinal outcomes—for comparison.

Despite the more value-focused context, the pilot reproduced the qualitative pattern of the main results: realistic threat remained the strongest predictor of hostile actions and hateful language, symbolic threat showed weaker effects, their interaction was negative, and intergroup contact reduced both behaviors. Attitudinal outcomes showed the complementary pattern expected in this value-laden setting—symbolic threat exerted the comparatively stronger influence on identity and bias. Effect sizes for identity were extremely small, consistent with the tightly scaffolded group-identification prompts used in the pilot. Despite these attenuated magnitudes, the relative ordering of effects matched the main study, indicating that the threat–response hierarchy is robust across contexts.

Table E58: Predicting hourly hostile action rates.

Predictor	β	SE	<i>p</i>
Intercept	-8.56	0.33	< .001
Hostile action rate (lag)	0.05	0.01	< .001
Intergroup contact rate (lag)	-0.32	0.06	< .001
Symbolic threat	0.05	0.04	.293
Realistic threat	0.11	0.04	.011
Time	-0.17	0.05	< .001
Symbolic \times Realistic threat	-0.13	0.04	.002

Table E59: Predicting hourly hateful-language rates.

Predictor	β	SE	<i>p</i>
Intercept	-7.06	0.26	< .001
Hateful language rate (lag)	0.04	0.02	.067
Intergroup contact rate (lag)	-0.02	0.05	.610
Symbolic threat	0.13	0.14	.340
Realistic threat	0.84	0.14	< .001
Time	-0.05	0.04	.222
Symbolic \times Realistic threat	-0.50	0.14	< .001

Table E60: Predicting attitude probes (hourly averages).

Panel A: Group Identity			
Predictor	β	SE	<i>p</i>
Intercept	-0.00	0.00	.957
Group Identity (lag)	0.99	0.00	< .001
Symbolic threat	0.00	0.00	.009
Realistic threat	0.00	0.00	.007
Time	0.00	0.00	.626
Symbolic \times Realistic threat	0.00	0.00	.008
Panel B: Group Bias			
Predictor	β	SE	<i>p</i>
Intercept	-0.00	0.03	.901
Group bias (lag)	0.23	0.01	< .001
Symbolic threat	0.24	0.02	< .001
Realistic threat	0.04	0.02	.047
Time	-0.03	0.01	< .001
Symbolic \times Realistic threat	0.15	0.02	< .001

E.3.2 Set 3: Exploratory manipulation of agents' moral values

Set 3 was conducted as part of a separate exploratory project on moral values in generative agents. It replicated the main minimal-group design (Set 1) while additionally varying agents' moral-value orientations through belief statements adapted from the Moral Foundations Questionnaire 2 [72], inducing either high binding–low individualizing or high individualizing–low binding profiles. Because moral foundations inherently concern group-based values, this manipulation partially overlaps conceptually with *symbolic* threat, making detailed interpretation within the present framework problematic. We therefore do not analyze moral moderation effects here but report the overall results obtained when fitting the same models on the aggregate data, providing an additional robustness check and demonstrating that the main threat patterns generalize across populations with different moral orientations. Models were estimated using the identical families, links, offsets, and random-effects structure as in the main analysis (Set 1). We report the three primary specifications—predicting hostile actions, hateful language, and attitudinal outcomes—for comparison.

Results reproduced the qualitative patterns observed in the main study. Realistic threat increased hostile actions and hateful language, symbolic threat had additional independent effects, and their interaction was again negative. At the attitudinal level, symbolic threat exerted the stronger influence on group identity, while being less dominant on group bias. Overall, the replication confirms that the observed threat-response hierarchy generalizes across populations with differing moral orientations.

Table E61: Predicting hourly hostile action rates.

Predictor	β	SE	p
Intercept	-9.64	0.40	< .001
Intergroup contact rate (lag)	-0.61	0.04	< .001
Hostile action rate (lag)	0.03	0.01	< .001
Symbolic threat	0.17	0.04	< .001
Realistic threat	0.28	0.04	< .001
Time	-0.12	0.03	< .001
Symbolic \times Realistic threat	-0.08	0.04	.025

Appendix F Reproducibility, transparency, and robustness

F.1 Preregistration

- **Registration status:** The study was not preregistered due to its initially exploratory nature, aimed at developing and stress-testing a novel generative-agent framework. Instead, robustness was established through multiple independent replication sets with identical threat manipulations and analytical pipeline.

Table E62: Predicting hateful-language rates.

Predictor	β	SE	<i>p</i>
Intercept	-7.10	0.23	< .001
Hateful language rate (lag)	0.01	0.01	.420
Intergroup contact rate (lag)	0.06	0.03	.027
Symbolic threat	0.49	0.07	< .001
Realistic threat	0.92	0.07	< .001
Time	0.03	0.03	.253
Symbolic \times Realistic threat	-0.33	0.07	< .001

Table E63: Predicting attitudinal outcomes.

Panel A: Group Identity			
Predictor	β	SE	<i>p</i>
Intercept	-0.01	0.03	.804
Group Identity (lag)	0.14	0.01	< .001
Symbolic threat	0.48	0.01	< .001
Realistic threat	0.16	0.01	< .001
Time	0.15	0.00	< .001
Symbolic \times Realistic threat	0.02	0.01	.024

Panel B: Group Bias			
Predictor	β	SE	<i>p</i>
Intercept	0.00	0.04	.932
Group bias (lag)	0.07	0.01	< .001
Symbolic threat	0.47	0.03	< .001
Realistic threat	0.27	0.03	< .001
Time	-0.02	0.00	< .001
Symbolic \times Realistic threat	0.11	0.02	< .001

- **Scope and replication pipeline:**

- *Set 0:* Preliminary pilot simulations using a non-minimal group paradigm designed to elicit stronger identity fusion and moral identity in agents. Although conceptually distinct from the main study, these runs yielded the same qualitative threat-response patterns and are reported in Section 5.3 of the Supplementary Materials for completeness. This pilot motivated the shift to the minimal-group paradigm used in the main experiments, to test whether similar dynamics also emerge from minimal group settings.
- *Set 1:* Main minimal-group design reported in the present manuscript.

- *Set 2*: Replication of the main design with added structural manipulations (segregation/integration \times equal/unequal group size) reported in the current manuscript.
- *Set 3*: Additional dataset from a separate project varying agents’ moral-value profiles (high binding–low individualizing versus high individualizing–low binding). Although conceptually distinct, analyses using this dataset reproduced the principal threat effects observed in the main study. Corresponding results are reported in SI Section 5.3.
- **Consistency of design and analysis:** Across all replication sets, the underlying threat manipulations, environment structure, and statistical modeling strategy were held constant. No alternative model specifications or post-hoc analytical decisions were introduced, providing a functionally equivalent safeguard to preregistration in terms of design transparency and analytical consistency.

F.2 Transparency and accessibility of materials

- **Repository and materials:** All code, prompt templates, configuration files, data-processing pipelines, and statistical analysis scripts are made publicly available at <https://osf.io/5ac3d>. The repository includes comprehensive setup instructions, version-controlled environment files (`environment.yml`, `requirements.txt`), and an automated installation script. These resources enable full replication of simulation orchestration, data aggregation, data processing (e.g., text classification, and extraction of agent probes), and subsequent statistical analyses and reporting.
- **Reproducibility testing:** The complete pipeline was independently tested on three distinct computing environments: (i) a local Linux (Ubuntu 24.04.2 LTS) workstation equipped with an NVIDIA RTX 5090 GPU, and (ii) a distributed high-performance computing cluster running Rocky Linux 8 NVIDIA A100 GPUs, iii) a commercial on-demand GPU cluster running NVIDIA RTX 4090 GPUs⁸. All configuration details (e.g., CUDA versions, driver specifications, installation templates) are documented in the repository for verification and reuse.

F.3 Dealing with nondeterminism

- **Reproducibility scope:** Given the inherent stochasticity of autoregressive language models, low-level trajectories (e.g., exact utterances or micro-actions) cannot be identically reproduced even with fixed random seeds. Instead, reproducibility is achieved at the level of analysis: rerunning the same code and experimental configuration should yield equivalent distributions of behaviors and attitudinal outcomes, that reproduce the same qualitative threat effects (e.g., main effects of realistic and symbolic threat). This strategy minimizes both bias and variance by aggregating across multiple independent, randomized realizations of the same experimental design that show variance and diversity on the micro-level but consistency on the macro level.
- **Inference parameters and controlled variability:** Generative reasoning and dialogue processes were executed with, e.g., `temperature = 0.8`, `top-p = 0.9`, and

⁸ www.vast.ai

`top-k` = 50 to promote creative but coherent agent behavior, following the settings used by Park et al. [10]. For deterministic subroutines such as spatial navigation, path planning, and action timing, temperature was reduced to 0.01 (minimal value) to ensure reproducible outputs and environment consistency.

- **Randomization of potential confounds:** Group assignments and agent-level attributes were re-randomized in each run to prevent systematic biases arising from fixed persona combinations. In robustness experiments (e.g., simulations varying agent moral profiles or demographic features), assignments of those features were also randomized per run. This approach both diversifies emergent social trajectories and ensures that findings are not driven by specific persona configurations.
- **Randomness control and replication:** Each simulation run used fixed random seeds for all stochastic components, including both model inference and procedural elements (`random`, `numpy` packages). For each experimental cell in the 2×2 design, we executed 10 independent runs using distinct seeds (`seed = N`, where $N \in [0, 9]$). Data from all runs were aggregated for analysis, with run identifiers preserved to allow hierarchical modeling of within- and between-run variance. This design minimizes the influence of any single idiosyncratic trajectory on the model results.

F.4 Model stability and accessibility

- **Model and inference engine:** All simulations used the quantized model version `matatonic/Mistral-Small-24B-Instruct-2501-6.5bpw-h8-ex12`, accessed on May 1st, 2026. The model weights were downloaded from HuggingFace⁹ and used unchanged across all inference runs. Inference was executed with the `ExLlamaV2` engine, consistent with the setup described in the Methods of the main text.
- **Model accessibility:** Both the model weights and inference engine are publicly available, enabling full replication of the simulation pipeline.
- **Model stability:** To ensure version stability, the downloaded weights were stored locally and used identically throughout all runs. The version of the inference engine used (`ExLlamaV2 v0.2.3`) is specified in the installation script and environment files to enable future replications to reproduce an identical inference setup. To further reduce randomness in generation, we ran all simulations on the same hardware (RTX 4090 GPUs).

F.5 Validation and justification

- **LLM output validation:** Because the present work involves open-ended, sequential simulations rather than isolated one-shot completions, validation cannot be performed at the level of individual prompt-response pairs. Instead, validation operates at the analytic level—whether aggregate behavioral patterns across runs reproduce theory-consistent threat effects. For the linguistic classifiers used in analysis, we report the model architecture, fine-tuning dataset, and validation performance. Specifically, moral-language classification employed a `roberta-base` model fine-tuned on the Moral Foundations Twitter Corpus [MFTC; 66], achieving a cross-validated F1 score of 0.76. Sentiment and hate-speech detection relied on two large

⁹<https://huggingface.com/matatonic/Mistral-Small-24B-Instruct-2501-6.5bpw-h8-ex12>

DeBERTaV3 models [73], fine-tuned on the `tweet_eval` dataset [74], which achieved validation F1 scores of 0.74 (sentiment) and 0.61 (hate). DeBERTa improves upon BERT and RoBERTa by introducing disentangled attention and enhanced mask decoding, yielding superior performance on most natural-language-understanding benchmarks.

- **Validation of experimental manipulations:** Because the primary outcomes emerge through extended agent behavior rather than single-response accuracy, validation focuses on verifying that manipulations produced the intended internal states. To confirm that the threat manipulations were effective and stable, we modeled agents' probed perceptions over time (see SI Section 5.1). Mixed-effects analyses showed that agents in symbolic- and realistic-threat conditions reported persistently elevated perceptions of the targeted threat type (near ceiling on a 1–7 scale) and baseline levels of the non-target type, with negligible temporal drift. These results confirm that the reinforcement–suppression procedure maintained stable, condition-appropriate threat representations throughout the simulation period.
- **Robustness of prompts and model settings:** The simulation framework employs dynamic prompt templates (all provided in the public repository) governing agent reasoning, planning, dialogue generation, and action execution. Templates were designed to enforce consistent output structure and valid action formats (e.g., proper location names or durations) while allowing open-ended reasoning. Prompt and inference configurations were optimized for stability in long-horizon trajectories rather than isolated completions, balancing computational efficiency, reproducibility, and behavioral realism. Manipulation-related prompts (e.g., group membership, threat perception, moral value orientation) were theory-driven and tightly specified to preserve construct validity; therefore, no prompt or parameter variation was conducted, as doing so may change the conceptual meaning of the experimental manipulation. Future work could explore prompt-level robustness among variations that are still conceptually aligned, but such analyses are beyond the present study's scope. Instead, as specified above, we validated the experimental manipulation by probing the agents as described above.
- **Validation level and robustness to nondeterminism:** Robustness was assessed through independent repetitions of each experimental condition, conducted with distinct random seeds and randomized agent assignments. Mixed-effects models were fitted across all runs, treating inter-run variability as a random source of noise—analogous to variability across independent replications in behavioral experiments—such that systematic effects would cancel out if driven purely by stochasticity. The persistence of strong and theory-consistent effects across the aggregated data therefore indicates that the observed patterns are robust to nondeterminism in LLM generation and initialization.
- **Interpretation limits and lack of human ground truth:** Direct ground-truth validation against human data is not feasible for this paradigm for two reasons. First, comparably fine-grained, longitudinal datasets capturing real-world intergroup dynamics at the level of individual behaviors and interactions do not exist, making one-to-one correspondence unattainable. Second, even available causal evidence

from laboratory or survey studies typically relies on self-reported attitudes or hypothetical judgments under low-stakes conditions, which differ fundamentally from the repeated, consequence-bearing interactions modeled here. The simulations therefore serve as a complementary framework for testing causal hypotheses about threat and social dynamics in a naturalistic yet fully controllable setting. To assess ecological plausibility, we also examined an independent hiring-bias scenario (SI Section 5.1) within the exact same experimental paradigm to test whether other related biases and discrimination that we can compare to human data emerge. Agents reproduced well-established human social-judgment patterns—physical attractiveness increased hiring success ($\hat{\beta} = 0.35, p < .001$) while a foreign accent decreased it ($\hat{\beta} = -0.26, p = .001$)—closely matching meta-analytic human effect sizes [70, 71] (preference for attractive candidates $d = .37$; preference for foreign accents $d = -.47$). These findings demonstrate that the same architecture generating intergroup threat dynamics also captures realistic, socially patterned decision biases. Nonetheless, all results should be interpreted as model-based generative approximations of social behavior rather than direct measurements of human populations.

F.6 Data processing and error handling

- **Pipeline transparency:** The full data-processing pipeline—from raw simulation logs to analytic datasets—is implemented in open scripts included in the public repository. These scripts document every transformation step, including log parsing, variable extraction, aggregation, and data merging. No manual data editing was performed.
- **Error handling and exclusions:** All logged agent actions, plans, and conversations were retained for analysis. The simulation framework was extensively tested prior to data collection to prevent invalid or malformed outputs (e.g., missing actions, non-executable plans, or incomplete dialogues) that could otherwise interrupt simulation continuity. Because the environment requires valid outputs for progression, such events are functionally precluded during runtime. Consequently, no data exclusions or post hoc error corrections were necessary.
- **Bias analysis:** While no data were excluded, pipeline diagnostics confirmed that logged data volume and composition were consistent across experimental conditions, reducing the likelihood of systematic bias related to simulation integrity or runtime variability.