

Article

Engineered Misunderstanding Under Psychological Warfare: A Bayesian Signaling Game of Felt-Understanding Collapse in the German Atomausstieg

Ryanne R. L. Fairchild

JFK School of Psychology and Social Sciences, National University, San Diego, CA 92123, USA;
r.fairchild7468@student.nu.edu

Abstract

Russian state-sponsored disinformation has been described in policy and the operational literature, but it is less often formalized in game-theoretic terms. Here, a two-layered formal model is developed showing how adversarial perturbation of a communication channel can collapse cross-group felt understanding—the third-order intentional state/belief structure, established empirically by Livingstone, in which one group believes its perspectives are recognized and accepted as valid by another. The Analytical Model is a static Bayesian signaling game with binary types and a noisy channel parameterized by perturbation rate π . The Analytical Model shows that when recognition benefits exceed signaling costs, there exists a perturbation threshold $\pi^* = 1 - c_R/(u_R \cdot p)$ above which mutual misrecognition becomes the unique Perfect Bayesian Equilibrium outcome. The Computational Model embeds this logic in an agent-based simulation on a homophilic stochastic block model and scale-free networks with continuous recognition capacity. Four substantive findings emerge: the closed-form analytical threshold from the Analytical Model predicts the boundary of collapse in the dynamic networked simulation; high network homophily protects cooperative behavior below π^* but provides no rescue above it; bridge seeding—the placement of recognition-capable agents at structurally central cross-group positions—is the most effective of three policy interventions tested, rescuing cooperation even above π^* ; and uniform adversarial volume is approximately as damaging as strategically targeted adversarial precision across both small dense and large scale-free topologies, qualifying the operational claim that targeted disinformation should strictly outperform volume-based approaches. The model is illustrated with the German Atomausstieg (nuclear phase-out) case, and implications for clinical psychology, public policy, and intergroup recognition under psychological warfare are discussed.

Keywords: signaling games; incomplete information; disinformation; felt understanding; agent-based modeling; Bayesian persuasion; network polarization; intergroup conflict



Academic Editor: Yasuko Kawahata

Received: 23 May 2026

Revised: 26 June 2026

Accepted: 26 June 2026

Published: 2 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Modern democracies face a class of threats that conventional political analysis is poorly equipped to describe. The weaponization of digital information networks by state-sponsored actors has intensified intergroup conflict in ways that cannot be reduced to the standard categories of propaganda, persuasion, or partisan polarization. What such operations do, in their most effective form, is not change anyone's mind. They make it impossible for groups to recognize one another as the kind of interlocutor with whom

dialogue is even possible. As recent psychohistorical analyses have demonstrated, these campaigns constitute a “war on reality,” where the strategic objective is not to win an argument but to transform technical policy questions into intractable symbols of identity and moral conflict (Fairchild, 2026).

The most salient operational realization of this threat is the Russian “firehose of falsehood” propaganda model, characterized by high-volume, multichannel, internally inconsistent messaging designed to saturate the information environment rather than to persuade through argument (Paul & Matthews, 2016). Recent scholarship has complicated this operational picture. Fridman (2023) argues that Russian information operations may be better understood as “strategy without design”—a wayfinding process driven by immediate situational opportunities rather than top-down navigation toward predefined goals—which, if correct, further explains the empirical preference for high-volume, low-precision operations over costly targeted campaigns. The psychological casualty of this relentless signal perturbation is felt understanding: the belief that members of an outgroup understand and accept the perspectives of ingroup members as authentic and subjectively valid, including those members’ beliefs, values, and self-definition (Livingstone et al., 2020). Across five studies in contexts including Scotland–UK relations, UK–EU relations, Catholic–Protestant relations in Northern Ireland, Basque–Spanish relations, and adolescent–adult intergenerational relations, Livingstone and colleagues established felt understanding as one of the strongest and most consistent predictors of intergroup trust, political separatism, and actual voting behavior, with standardized regression coefficients exceeding 0.4 even after controlling for outgroup beliefs, metabeliefs, and beliefs about appearing prejudiced. When the signals supporting felt understanding are degraded, trust and cooperation become difficult to sustain.

The German Atomausstieg (nuclear phase-out) illustrates the consequences. State-sponsored disinformation, operating through Gazprom-linked NGO infrastructure and amplified through aligned media, systematically exploited cognitive vulnerabilities—specifically the “affect heuristic” and “solution aversion” (Campbell & Kay, 2014; Slovic et al., 2007)—to paralyze rational climate policy (Fairchild, 2026). By amplifying the “dread risk” associated with nuclear energy (Chilton et al., 2006), these operations successfully fractured cross-group dialogue between environmentalists and pro-nuclear publics. The downstream consequences included accelerated fossil dependence, an estimated 800 excess deaths per year from coal-fired air pollution that replaced nuclear generation (Jarvis et al., 2022), and a geopolitical vulnerability that constrained Germany’s response to the 2022 invasion of Ukraine exemplified by Berlin drawing scorn from Ukrainian officials and European partners after offering only 5000 helmets as military aid amid fears of Russian retaliation and energy reprisals (Boffey & Oltermann, 2022).

In this respect, Atomausstieg is useful as a theoretically rich case in which affective risk perception, identity-protective cognition, NGO/proxy influence, and strategic information flooding converged. The ultimate psychological objective of such operations is to exhaust populations to the point where they withdraw from engagement altogether, seeing truth as too elusive to pursue (Fairchild, 2026; Lewandowsky et al., 2017; Pomerantsev & Weiss, 2014). This aligns with the mechanics of fifth-generation warfare (5GW), which operates at the psychological and societal level; as Krishnan (2024) states, “the main battlespace of 5GW is the human mind” (p. 46). This relentless barrage of disinformation drives the public either into “learned helplessness,” where systemic action feels impossible, or into an “illusion of control” through materially ineffective personal behaviors (Fairchild, 2026; Langer, 1975; Pomerantsev & Weiss, 2014; Seligman, 1975). Despite the empirical reality of these campaigns and the centrality of the felt-understanding construct, no formal account yet explains how induced exhaustion degrades felt understanding at equilibrium in the

Atomausstieg context—that is, how short-term signal perturbation by a hostile third party translates into long-run misrecognition that destabilizes democratic publics. Consequently, the clinical psychology significance of engineered misunderstanding should not be framed only as a problem of belief correction or political persuasion. Trauma exposure, PTSD, and depression are associated with disruptions in autobiographical memory, and Jobson et al. (2023) note that the relationship between psychopathology and misinformation effects remains underdeveloped, especially in clinical populations.

A two-layered model is developed. The Analytical Model is a static Bayesian signaling game in which two domestic groups send costly signals to one another through a channel perturbed by an adversary at rate π . The Analytical Model posits that there exists a threshold $\pi^* = 1 - c_R / (u_R \cdot p)$, depending only on the cost of authentic signaling c_R , the payoff to mutual recognition u_R , and the prior proportion p of recognition-capable agents, above which the unique Perfect Bayesian Equilibrium is mutual misrecognition. The Computational Model embeds this logic in an agent-based simulation on a homophilic stochastic block model and scale-free networks. Two substantive results follow that exceed the scope of the Analytical Model. First, high within-group homophily protects cooperative behavior below the threshold but provides no rescue above it: the network structure that stabilizes dialogue under low perturbation becomes an accelerator of within-group radicalization once the threshold is crossed. Second, across both small dense and large scale-free network topologies, uniform adversarial volume and strategically targeted adversarial precision produce approximately equivalent damage to cross-group recognition. The model qualifies a substantive prediction of the firehose-of-falsehood literature: strategic precision does not outperform equivalent-budget volume in the topologies tested, though it also does not catastrophically underperform. An adversary indifferent on the substantive margin will rationally prefer the operational approach with lower intelligence and identification overhead, which is consistent with the empirical preference for volume in state-sponsored disinformation operations.

Section 2 reviews the three literature areas the analysis bridges: the social psychology of intergroup felt understanding, the game-theoretic analysis of strategic communication and Bayesian persuasion, and the agent-based modeling of opinion dynamics and disinformation. Section 3 develops the Analytical Model and proves the central proposition. Section 4 specifies the Computational Model and bridges the analytical and simulation framings. Section 5 reports the simulation results across four substantive figures and one robustness check. Section 6 discusses implications for the Atomausstieg case and for the design of recognition-restoring interventions and sets out the limitations of the analysis. Section 7 concludes with directions for further work.

2. Literature Review

This research bridges three scholarly domains: the social psychology of intergroup recognition, the game-theoretic analysis of strategic information transmission, and the computational modeling of network polarization. Each in turn is reviewed, emphasizing the gaps the present analysis addresses.

2.1. The Psychology of Felt Understanding

Intergroup-relations research has long studied the role of beliefs about outgroups (prejudiced attitudes, stereotypes, threat perceptions) and, more recently, metabeliefs about what an outgroup believes about the ingroup (e.g., Frey & Tropp, 2006; Vorauer et al., 1998). Felt understanding extends this line of work by introducing a higher-order perspective: the belief about an outgroup's belief about the ingroup's own perspectives, beliefs, and values (Livingstone et al., 2020). It is third-order intentional in the sense of Dennett (1987)

and involves a second-order theory of mind as in [Kinderman et al. \(1998\)](#): a belief about a belief about a belief. The construct draws conceptually on interpersonal work, emphasizing feeling understood as fundamental to relational well-being ([Reis et al., 2017](#); [Rogers, 1995](#)) and refines it for intergroup contexts in which perceptions occur through the lens of salient social identities.

The empirical case for treating felt understanding as a primary construct rather than a derivative of better-studied variables is strong. Confirmatory factor analyses across three studies (combined $N > 7000$) preferred a model in which felt understanding loaded as a distinct factor from outgroup beliefs, metabeliefs, and beliefs about appearing prejudiced, with a normalized Akaike weight of 0.999 ([Livingstone et al., 2020](#)). In predictive models, felt understanding emerged as one of the strongest predictors of institutional trust ($\beta = 0.63$ in Scotland–UK; $\beta = 0.52$ in UK–EU), of support for political separatism ($\beta = -0.48$), and of actual Brexit voting behavior ($\beta = -0.43$). Study 5 separated felt understanding from felt liking experimentally, demonstrating that feeling understood predicts positive intergroup orientations even when subjects are told the outgroup dislikes them. This predictive structure has since replicated beyond the original Western samples: examining Japanese–Chinese relations in Japan, [Ioku and Watamura \(2025\)](#) found that felt understanding uniquely predicted core intergroup outcomes—outgroup trust, intergroup orientation, and several action intentions—even after controlling for outgroup beliefs, metabeliefs, and beliefs about appearing prejudiced, consistent with the construct’s grounding in a culturally general second-order theory of mind.

Because felt understanding depends on the accurate interpretation of observable signals sent by the outgroup, it is uniquely vulnerable to environments in which communication channels are intentionally distorted. This vulnerability has not previously been formalized.

2.2. Strategic Information Transmission and Bayesian Persuasion

The game-theoretic literature on strategic communication originates with the cheap-talk framework of [Crawford and Sobel \(1982\)](#), which characterizes how divergent preferences between sender and receiver limit the amount of information that can be transmitted in equilibrium. In contexts where signaling is costly, and types satisfy the single-crossing property, separating equilibria can convey hidden types, with off-path beliefs restricted by refinements including the Intuitive Criterion ([Cho & Kreps, 1987](#)).

More recently, the literature has shifted toward Bayesian persuasion, in which a sender commits to an information disclosure policy that influences the actions of a rational Bayesian receiver ([Kamenica & Gentzkow, 2011](#)). Recent work extends this framework to audiences that are not uniformly rational—for example, persuasion of both sincere and strategic voters ([Kerman et al., 2024](#)) and information design that remains robust when receivers respond through a quantal-response rule rather than exact best response ([Feng et al., 2024](#)); the latter connects directly to the present model, whose agents select actions by exactly such a rule. This framework has been productively applied to intergroup conflict: [Hennigs \(2021\)](#) shows that a mediator who commits to an information disclosure policy about the military strength of conflicting parties can reduce the ex ante probability of war. The present analysis inverts this logic. Rather than a benevolent mediator providing information to prevent conflict, a hostile third party is modeled as degrading information to create misrecognition. The adversary’s instrument is not disclosure but perturbation: a policy that determines the probability with which signals sent between two domestic groups are corrupted in transit. The closest formal precedent to the present model is [Little \(2017\)](#), who develops a signaling game in which a government chooses a level of propaganda that inflates a public signal of its performance, observed by a population containing

both credulous and non-credulous citizens with a coordination motive. Little (2017) demonstrates that when the coordination motive dominates, non-credulous citizens mimic the behavior of the credulous, so that propaganda shapes mass behavior even without shaping mass beliefs. The present analysis furthers this logic from a single-group setting, in which coordination occurs among citizens vis-à-vis their government, to a two-group setting, in which the equilibrium object is mutual recognition between domestic publics rather than support for a regime. Where Little's (2017) adversary chooses manipulation intensity to influence a single audience, the adversary modeled here chooses a perturbation rate on signals exchanged between two groups, and the equilibrium failure mode is misrecognition rather than misplaced support. This extends the smaller literature on strategic communication over noisy channels (Le Treust & Tomala, 2016), with the noise re-interpreted as an adversarial policy variable rather than a natural feature of the environment.

On the defensive side, Balcaen et al. (2021) develop a game-theoretic deterrence model in which a defender allocates resources across conventional and hybrid domains against a revisionist challenger, incorporating prospect-theoretic insights to explain why defenders systematically under-invest in hybrid deterrence. Their model addresses the allocative problem facing the targeted state but does not characterize how perturbation operates on the social-psychological substrate of the targeted population itself. The present model is complementary: it takes the perturbation as given and asks how it propagates through intergroup recognition dynamics to produce the misrecognition equilibrium that, in turn, makes the defender's allocative problem so difficult to solve. More broadly, the contest between an information-degrading adversary and a defending target situates the present model within the attacker-defender game tradition, recently surveyed by Hausken et al. (2024), in which signaling, deception, and incomplete information are recurring features of the strategic interaction.

2.3. Disinformation Dynamics and Network Polarization

The operational characterization of adversarial perturbation that anchors this analysis is the "firehose of falsehood" model, which describes Russian state-sponsored propaganda as high-volume, multichannel, rapid, repetitive, and lacking commitment to internal consistency (Paul & Matthews, 2016). Fridman (2023) further argues that Russian information operations exhibit "inherent strategic ambiguity" between objectives and actions, with the same operational activity simultaneously serving tactical, operational, and strategic informational ends across multiple target audiences. This characterization is consistent with the volume-over-precision empirical pattern that the present model's computational results help to rationalize. U.S. Department of State (2020) and U.S. House of Representatives, Committee on Science, Space, and Technology (2018) documentation establishes that this approach has been deployed against Western energy policy debates specifically, including through covert support of environmental NGOs opposing fracking and nuclear power (Harvey, 2014; Bennhold & Solomon, 2022). Bradshaw and Howard (2019) characterize Russia as a "high capacity" actor in their comparative inventory of organized social media manipulation.

The effects of such interventions depend on the topological structure of the target population. Recent agent-based and dynamical models of opinion formation show that homophily and reinforcing radicalization drive moderate populations toward persistent, polarized echo chambers (Baumann et al., 2020). This sociophysics tradition is surveyed comprehensively by Tsintsaris et al. (2024); recent contributions extend it to signed networks with antagonistic as well as cooperative ties (Vendeville & Diaz-Diaz, 2025) and show empirically that network-based and language-based measures of polarization can diverge (Durrheim & Schuld, 2025). The introduction of inauthentic agents into such networks

intensifies polarization: [Lu and Lee \(2025\)](#) demonstrate, in a bounded-confidence model with symmetrically deployed political bots, that adversarial agents increase opinion polarization across a range of network topologies. Complementing these dynamical accounts, [Acemoglu et al. \(2024\)](#) provide an equilibrium model of how online misinformation is shared and propagates across a social network. The Computational Model of the present analysis expands in scope this computational literature by grounding agent update rules in the Bayesian signaling thresholds derived in the Analytical Model, enabling a test of the comparative effectiveness of targeted versus volume-based adversarial strategies and the comparative effectiveness of three classes of intervention against them.

3. The Analytical Model: A Bayesian Signaling Game

The Analytical Model represents two domestic identity groups exchanging costly signals through a channel whose fidelity is degraded by an adversarial third party. The model is static and uses a binary type space, which yields the cleanest characterization of the equilibrium structure. Section 4 relaxes both restrictions in the dynamic agent-based extension.

3.1. Players, Types, and Signals

There are two domestic identity groups, A and B, and one adversary, R. Each group is represented by a single agent in the formal model; this is the standard representative agent reduction used throughout the signaling game literature ([Spence, 1973](#); [Cho & Kreps, 1987](#)), in which the equilibrium analysis describes the modal behavior of the underlying population. The adversary R is non-strategic: R is modeled as a parameter of the environment that determines the perturbation policy applied to the signal channel. The implications and limits of this simplification are discussed in Sections 3.6 and 6.4.

Each agent has a private type drawn from a binary space:

$$\theta_i \in \Theta = \{R, N\}, \quad (1)$$

where R denotes “recognition-capable” and N denotes “recognition-incapable/non-recognition-oriented agent.” A recognition-capable agent, on observing an authentic signal of the other group’s perspectives, updates toward recognition of those perspectives as subjectively valid in the sense of [Livingstone et al. \(2020\)](#). A recognition-incapable agent is one for whom recognition would constitute an identity threat ([Livingstone et al., 2009](#)) and who, therefore, produces a dismissive response regardless of signal content. Types are drawn independently from a common prior $P(\theta = R) = p$ with $p \in (0, 1)$, assumed common knowledge. The binary structure is a deliberate simplification of the continuous recognition capacity that the agent-based extension models as a simplifying abstraction; the empirical scale itself measures receiver-side felt understanding rather than objective recognition capacity; the agent-based extension in Section 4 uses $\theta \in [0, 1]$ drawn from a Beta distribution.

Each agent chooses an action $a \in A = \{a_R, a_D\}$, interpretable as recognition (a_R) or dismissal (a_D) of the other group’s perspectives. The action determines a signal $s \in S = \{s_A, s_D\}$ sent through the channel to the other group: a_R generates s_A (an “authentic” signal revealing the agent’s perspectives in a form that, if received undistorted, licenses recognition of them as genuine), and a_D generates s_D (a “dismissive” signal conveying neither perspective nor openness to recognition). Agents choose actions without observing the other’s action; payoffs depend on what the other agent observes through the perturbed channel, as specified below.

A modeling note. [Livingstone et al.’s \(2020\)](#) 15-item scale measures receiver-side judgments—ingroup members’ beliefs about whether outgroup behavior conveys recognition. The model reconstructs the sender side that the receiver-side measure presupposes.

This reconstruction is defensible but is a modeling assumption rather than a direct empirical mapping, revisiting the bridge between sender-side signal and receiver-side judgment in Section 4. Accordingly, simulation outputs are interpreted as observable recognition signal rates from which felt understanding could be inferred, not as direct scores on the Livingstone et al. (2020) scale.

3.2. Costs and Payoffs

Signaling is costly, with cost depending on type:

$$c(s_A | \theta = R) = c_R, c(s_A | \theta = N) = c_N, c_N > c_R > 0, \quad (2)$$

with $c(s_D | \theta) = 0$ for both types. This is the standard single-crossing structure of separating signaling games (Spence, 1973). Substantively, a recognition-capable agent finds authentic signaling less costly because it is congruent with their identity stance (willing vulnerability, articulation of values, openness to disagreement); a recognition-incapable agent finds it more costly because authentic signaling is incongruent with the identity-protective orientation that predicts dismissive intergroup behavior regardless of signal content (Livingstone et al., 2009), requiring either suppression of those dismissive responses or fabrication of signals the agent does not actually endorse.

For agent i in group A facing agent j in group B, with i 's signal arriving at j as $\hat{s}_i$ after channel perturbation, the payoff is:

$$u_i = u_R \cdot \mathbb{1}\{\hat{s}_i = s_A, \theta_j = R\} - u_I \cdot \mathbb{1}\{a_i = a_R, \theta_i = N\} - c(s_i | \theta_i). \quad (3)$$

The first term, $u_R > 0$, is the recognition payoff: agent i is rewarded when its signal arrives at j as authentic ($\hat{s}_i = s_A$) and j is of a type capable of recognizing it ($\theta_j = R$). Both conjuncts are required: a garbled signal is not recognized even by a capable receiver, and an authentic signal reaching an incapable receiver yields no recognition. This formalizes the empirical regularity that felt understanding requires both that perspective signals be transmitted faithfully and that the receiver be capable of accepting them as subjectively valid (Livingstone et al., 2020). The second term, $u_I > 0$, is an identity-protection penalty: a recognition-incapable agent who plays in the recognition mode ($a_i = a_R$) suffers an identity-threat cost, formalizing the Livingstone et al. (2009) finding that perceived identity threat predicts defensive intergroup orientations. The third term is the signaling cost defined above.

The following magnitude conditions on the payoff parameters are assumed:

- $u_R > c_R$: a recognition-capable agent prefers authentic signaling when transmission and reception both succeed;
- $u_R < c_N$: a recognition-incapable agent's authentic-signaling cost exceeds the recognition payoff even when transmission and reception succeed;
- $u_I + c_N > u_R$: the combined identity and signaling cost makes recognition-mode play strictly dominated for type N, regardless of beliefs about the channel or the other group.

These magnitude conditions guarantee the single-crossing property and are the foundation of the threshold result that follows.

3.3. The Adversarial Channel

Signals do not pass directly between groups; they pass through a channel whose fidelity is controlled by the adversary. Let $s \in S$ denote the signal an agent sends and $\hat{s} \in$

S denote the signal the receiver observes. The channel is characterized by a perturbation parameter $\pi \in [0, 1/2]$:

$$P(\hat{s} = s | s \text{ sent}, \pi) = 1 - \pi, P(\hat{s} \neq s | s \text{ sent}, \pi) = \pi. \quad (4)$$

When $\pi = 0$, the channel is undistorted; when $\pi = 1/2$, it is maximally distorted, and the observed signal is independent of the sent signal. Intermediate values represent the strength of the adversary's perturbation policy. Crucially, for any $\pi \in (0, 1/2]$, every observable signal is reached with positive probability under any sender strategy. This full-support property is of importance for the uniqueness argument in Section 3.4: it implies that no equilibrium refinement is required, because Bayes' rule pins down all posteriors directly.

The substantive interpretation of π differs from the noise terms that appear in conventional communication-theoretic signaling models (Le Treust & Tomala, 2016). Here, π is not a feature of the natural environment; it is the policy variable of an adversary that engineers, amplifies, or suppresses signals so as to degrade the relationship between what one group does and what the other group observes. In the Russian operational context, this corresponds to amplification of dismissive or hostile signals via inauthentic accounts and aligned media, injection of fabricated dismissive signals that no actual in-group member produced, and volume-based suppression of authentic signals through the firehose-of-falsehood approach (Paul & Matthews, 2016). The parameter π abstracts over these operational details and captures their aggregate effect on signal fidelity.

3.4. Equilibrium Analysis

A Perfect Bayesian Equilibrium (PBE) is used. A PBE consists of a strategy profile and a belief system such that strategies are sequentially rational given beliefs and beliefs are derived from strategies via Bayes' rule wherever the rule applies. The full-support property of the channel (Section 3.3) ensures that Bayes' rule applies everywhere; no off-path refinement is needed.

Felt understanding for group A about group B in this framework is the equilibrium posterior $\mu_A(\theta_B = R | \hat{s}_B)$ averaged over the equilibrium distribution of observed signals $\hat{s}_B$. This is the quantity that Livingstone et al.'s (2020) empirical scale measures via receiver judgments, and it is measured symmetrically for FU_B .

3.4.1. Benchmark: $\pi = 0$ (No Adversary)

When the channel is undistorted, the standard separating equilibrium exists under the magnitude conditions of Section 3.2. Type R plays a_R (sending s_A), type N plays a_D (sending s_D), and posteriors upon observing the corresponding signal are $\mu(\theta = R | s_A) = 1$ and $\mu(\theta = R | s_D) = 0$. Felt understanding in this equilibrium is $FU_A = FU_B = p$: well-calibrated to the underlying distribution of recognition capacity. This is the benchmark against which adversarial perturbation will be measured.

3.4.2. The Perturbed Channel: $0 < \pi < 1/2$

When the channel is perturbed, and the separating profile is being played, Bayes' rule gives:

$$\mu(\theta = R | \hat{s} = s_A) = (1 - \pi) p / [(1 - \pi) p + \pi (1 - p)], \quad (5)$$

and analogously for $\hat{s} = s_D$. These posteriors are well-defined and monotone in π : as π increases, the informativeness of the signal degrades, and at $\pi = 1/2$, both posteriors equal the prior p , conveying no information about type.

The substantive question is whether the separating profile remains a best response under the perturbed channel. Consider a recognition-capable agent's expected payoff from

playing a_R , given that the opposing population plays the separating strategy: the agent sends s_A ; with probability $(1 - \pi)$ the channel transmits the signal as $\hat{s} = s_A$; conditional on transmission, the probability that the receiver is type R is p ; the recognition payoff u_R is realized; and the cost c_R is paid with certainty. The expected utility is:

$$EU_R(a_R) = (1 - \pi) \cdot u_R \cdot p - c_R. \quad (6)$$

The expected utility from playing a_D is zero. Separating play, therefore, remains a best response for type R if and only if $(1 - \pi) \cdot u_R \cdot p > c_R$, or equivalently:

$$\pi < 1 - c_R / (u_R \cdot p) \equiv \pi^*. \quad (7)$$

When $\pi > \pi^*$, type R strictly prefers a_D , even though type R would prefer mutual recognition. The separating equilibrium unravels, and only the pooling equilibrium on a_D survives. At the boundary point $\pi = \pi^*$, type R is indifferent between a_R and a_D ; both the separating and pooling profiles are equilibria here, but the pooling profile is selected by the standard tie-breaking convention favoring the costless action a_D . The collapse result, therefore, extends to $\pi \geq \pi^*$. Felt understanding collapses to the uninformative prior, and population behavior is uniformly dismissive regardless of the underlying distribution of recognition capacity.

Proposition 1 (Felt-Understanding Collapse). *For any prior $p \in (0, 1)$ and cost structure satisfying the magnitude conditions of Section 3.2, there exists a threshold perturbation level $\pi^* = \max\{0, 1 - c_R / (u_R \cdot p)\}$ such that for all $\pi > \pi^*$, the unique Perfect Bayesian Equilibrium is the pooling equilibrium in which all agents of both groups play a_D . At $\pi = \pi^*$, type R is indifferent between a_R and a_D , and the pooling profile is selected under the standard tie-breaking convention, favoring the costless action. In this equilibrium, observed signals carry no information about types, and equilibrium felt understanding equals p while population behavior is uniformly dismissive regardless of the underlying recognition capacity distribution.*

Proof. The proof proceeds in three steps; complete algebra appears in Appendix A.

Step 1. Type N strictly prefers a_D in every equilibrium. By the magnitude condition $u_I + c_N > u_R$, even when the other group is certain to recognize, type N's payoff from a_R is at most $u_R - u_I - c_N < 0$, while a_D yields 0. So type N plays a_D regardless of π or beliefs about the other group. $\square$

Step 2. Given Step 1, type R plays a_R iff $(1 - \pi) \cdot u_R \cdot p > c_R$ against an opposing population playing the separating strategy. The expected utility calculation in Equation (6) and the threshold condition (7) follow directly.

Step 3. For $\pi > \pi^*$, the pooling equilibrium on a_D is the unique Perfect Bayesian Equilibrium. Because $\pi > 0$ gives the channel full support, every observable signal is reached with positive probability under any strategy profile; Bayes' rule, therefore, pins down all posteriors uniquely. Under the pooling profile, the observed signal carries no information about type, and the posterior on R equals the prior p . A type R deviation to a_R pays c_R and at best yields expected recognition payoff $(1 - \pi) \cdot u_R \cdot p < c_R$ by the strict threshold condition. The deviation is strictly unprofitable. At the boundary $\pi = \pi^*$, the deviation is exactly break-even, and both the separating and pooling profiles are equilibria; the pooling profile is selected under the standard tie-breaking convention favoring the costless action a_D . No off-path beliefs need to be specified or refined; the standard PBE suffices for uniqueness above the threshold and for selection at the boundary.

3.5. Comparative Statics

Three features of Proposition 1 are central to the paper's substantive interpretation.

First, the collapse is discontinuous in π at π^* . Below the threshold, separating play and calibrated felt understanding are sustainable; above the threshold, both collapse together. Small additional increments of adversarial capability that cross the threshold produce qualitative changes in equilibrium behavior. This matches the empirical pattern documented in the Atomausstieg case: a gradual buildup of disinformation infrastructure that produces, at a certain point, a rapid and durable collapse of cross-group dialogue (Fairchild, 2026).

Second, the threshold depends on the ratio $c_R/(u_R \cdot p)$. Interventions that reduce the cost of authentic signaling (lower c_R), increase the payoff to mutual recognition (higher u_R), or raise the proportion of recognition-capable agents (higher p) all raise the threshold and make collapse harder to engineer. This comparative statics structure generates the intervention taxonomy tested in Section 4 and corresponds substantively to the three classes of recognition-restoring intervention discussed in the clinical psychology implications of Section 6.

Third, the dead zone in which $\pi^* = 0$ is substantive. When the prior recognition capacity falls below c_R/u_R , the system is in the pooling basin regardless of adversarial activity: the expected reciprocation rate is too low to justify the signaling cost even on a clean channel. This is the formal counterpart to the empirical observation that some intergroup contexts—active violent conflict, deep historical grievance—are not amenable to dialogue interventions at all. Adversarial operations are most efficient at the margin: against populations whose recognition capacity is high enough to make dialogue possible in principle but not so high as to be invulnerable.

Figure 1 illustrates the phase geometry, with the stylized Atomausstieg parameterization ($p = 0.5$, $c_R = 0.35$, $u_R = 1.0$, yielding $\pi^* = 0.30$) marked. These magnitudes constitute a stylized, illustrative parameter set rather than an empirical calibration to German case data. The prior $p = 0.5$ follows from a deliberately uninformative, symmetric Beta(2, 2) over recognition capacity: neither group is assumed more recognition-capable than the other, and no estimate of the true population split is claimed. This symmetric choice is the least-biased maximum entropy prior in the sense of Jaynes (1957): absent ex ante grounds for either group to presume the other is predominantly willing or unwilling to recognize its position—the condition of mutual uncertainty characteristic of polarized but not yet hardened intergroup relations (Livingstone et al., 2020)—assigning $p = 0.5$ imports no directional assumption into the model, consistent with the standard rationale for constructing reference priors by formal rule (Kass & Wasserman, 1996); it is also the conservative choice, placing the prior at maximum uncertainty so that the perturbation dynamics, rather than an assumed asymmetry, drive the equilibrium outcome. The recognition payoff is normalized to unity, and the authentic-signaling cost is chosen so that its ratio to that payoff places the system at the empirically interesting margin identified above ($\pi^* = 0.30$), where cooperation is sustainable on a clean channel but collapses under moderate perturbation. The substantive conclusions do not depend on this particular set; a sensitivity analysis varying the prior, the cost, and the payoff is reported in Section 5.4 and Supplementary Figure S1.

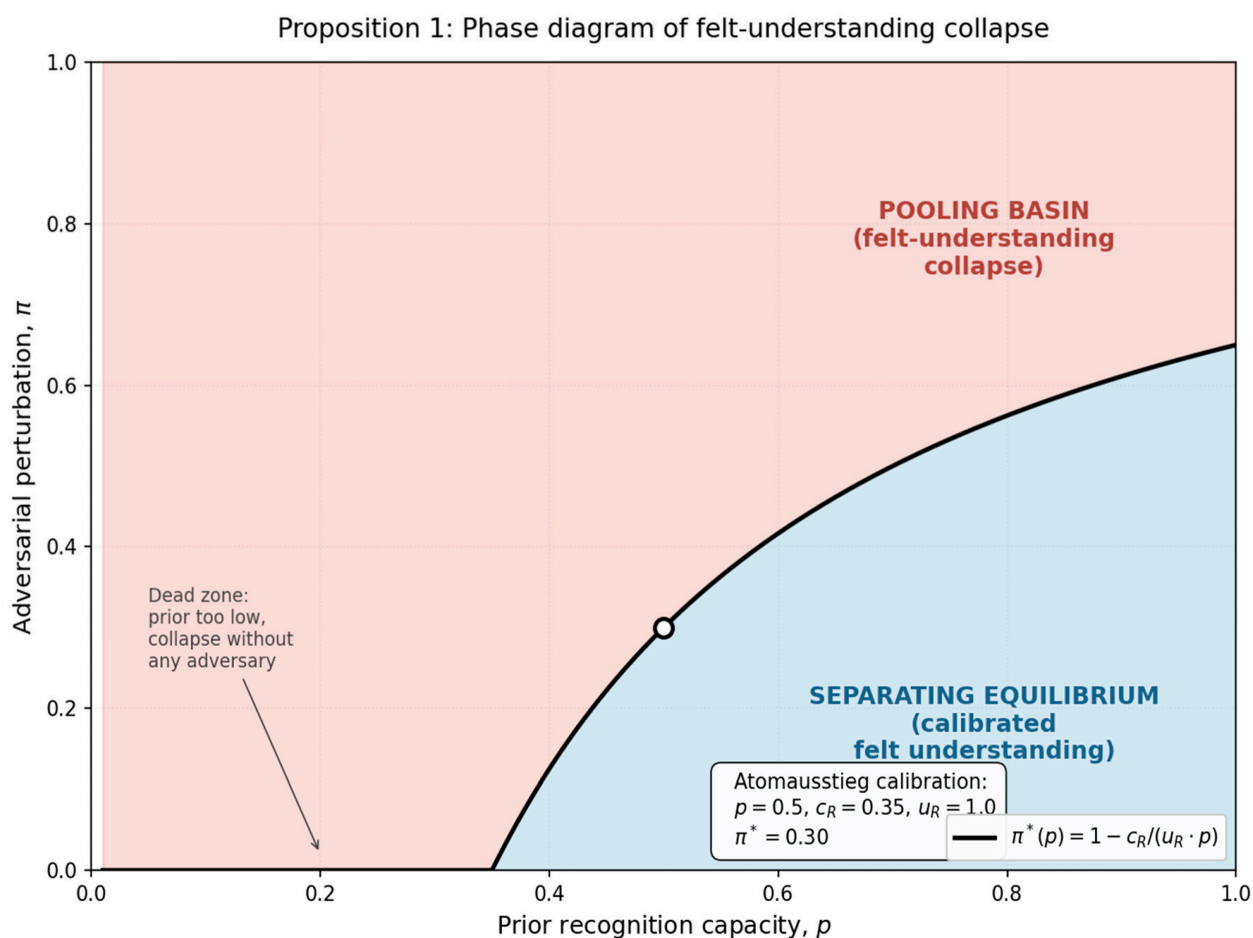


Figure 1. Phase diagram of Proposition 1. The threshold curve $\pi^*(p) = 1 - c_R/(u_R \cdot p)$ separates parameter combinations sustaining the separating equilibrium (lower-right region) from those in which only the pooling equilibrium survives (upper-left region). The stylized Atomausstieg parameterization ($p = 0.5, c_R = 0.35, u_R = 1.0$) yields $\pi^* = 0.30$, marked at the labeled point. The dead zone at low p , where $\pi^* = 0$ because $c_R/u_R \geq p$, indicates populations in which pooling obtains even without adversarial perturbation.

3.6. Modeling the Adversary: Justification of the Parametric Specification

The decision to model the adversary as a parameter of the environment rather than as a third optimizing player is deliberate and is defensible on three grounds. First, the reduction is the standard Stackelberg leader treatment. State-sponsored perturbation is a first-mover commitment: the adversary establishes a disinformation infrastructure—the volume, channels, and targeting of injected signals—to which domestic publics subsequently best respond. When one player commits to a policy before the others act, that policy enters the followers’ problem parametrically, and the followers’ equilibrium is properly analyzed for each fixed leader action. This is precisely the structure of the Bayesian-persuasion models on which the present analysis builds (Kamenica & Gentzkow, 2011), in which a sender commits to an information structure and Bayesian receivers best respond; the perturbation rate π is the analog of the sender’s committed information policy, here turned to the adversarial purpose of degrading rather than disclosing signal fidelity.

Second, the adversary’s problem is to choose a perturbation policy—a perturbation rate π , an injection budget γ , and a targeting policy τ —so as to maximize the long-run cross-group misrecognition D it induces, subject to a resource constraint:

$$(\pi^A, \gamma^A, \tau^A) \in \arg \max_{\pi, \gamma, \tau} D(\pi, \gamma, \tau), \text{ subject to } C(\pi, \gamma, \tau) \leq B. \quad (8)$$

where $D(\pi, \gamma, \tau) \equiv 1 - R(\pi, \gamma, \tau)$ is the induced misrecognition, R is the long-run cross-group action rate averaged over the two groups, C is the adversary's operational cost, and B is its resource budget. The objects derived above are precisely this maximand's structure over the adversary's strategy space. Proposition 1 establishes that D is a step function in π with its discontinuity at the threshold π^* ; the targeting experiments of Section 5.5 evaluate the dependence of D on τ at a fixed injection budget, finding it approximately flat and, where detectable at all, weakly favoring uniform injection. Appendix B states this targeting problem formally and shows that its solution is (weakly) uniform targeting on the topologies tested. A fully strategic adversary chooses the argmax of D over this surface; the present analysis supplies the surface.

Third, the parametric specification is arguably more faithful to the empirical adversary than a unitary rational optimizer specification would be. Fridman's (2023) characterization of Russian information operations as "strategy without design"—a wayfinding process driven by situational opportunity rather than top-down optimization toward a predefined objective—implies that the operational adversary may not solve a clean optimization problem at all. Treating perturbation capability as a policy parameter swept across its feasible range, and asking how the targeted social-psychological substrate responds at each level, brackets the adversary's internal decision process precisely where the empirical literature suggests that process is least well described by expected-utility maximization. This is also the standard division of labor in the attacker–defender tradition to which the analysis connects (Hausken et al., 2024; Balcaen et al., 2021): models that characterize the target's response to a given adversarial capability are logically prior to, and separable from, models that endogenize the adversary's choice of that capability. The fully endogenous three-player game—in which the adversary's argmax over D is solved jointly with a defender's allocation of the interventions analyzed in Section 5.4—is the natural next step, set out as such in Section 6.4; what the present analysis establishes is that the central inputs to that game, the shape of D in π and its flatness in τ , are already characterized here.

The Analytical Model provides the formal foundation for the analysis: a clean threshold result with closed-form comparative statics. What it does not capture are the dynamics by which felt understanding degrades over time, the network structure of cross-group signal transmission, the heterogeneity of agents within each group, or the differential effectiveness of intervention strategies. The agent-based extension in the next section addresses each.

4. The Computational Model: An Agent-Based Extension

The Analytical Model establishes that adversarial perturbation can collapse a separating equilibrium and the calibrated recognition-signal proxy associated with it. Three features of the empirical case that motivate the analysis are not captured by the Analytical Model and motivate the agent-based extension developed here. First, the Atomausstieg collapse unfolded over more than a decade; the Analytical Model is static and cannot accurately model the trajectory of degradation or the timing of intervention. Second, empirical disinformation operations can exploit network structure: target central nodes span weakly connected clusters and concentrate in homophilic communities. The representative agent reduction in the Analytical Model averages over this structure. Third, real agents do not sort cleanly into recognition-capable and recognition-incapable types; the Livingstone scale is continuous (felt-understanding construct is measured as a graded psychological judgment), and the cost–benefit calculation that determines whether an agent engages in cross-group recognition is heterogeneous across the population. The Computational Model addresses all three.

The simulation is constructed to be the dynamic, networked, finite-population analog of the Analytical Model. Specifically, in the limit of a large population, single round, no network structure, binary types, and symmetric parameters, the Computational Model reproduces the Analytical Model's equilibrium predictions. This reduction is verified in the validation suite reported in Section 5. The two layers, therefore, constitute a single integrated model rather than two separate analyses.

4.1. Network Structure

Agents are embedded in a graph $G = (V, E)$ with $V = N_A \cup N_B$ and $|N_A| + |N_B| = N$. Cross-group edge density is controlled by a homophily parameter $h \in [0, 1]$, where $h = 0$ corresponds to a random graph ignoring group membership and $h = 1$ to complete intragroup connection with zero intergroup connection. Empirically realistic values for polarized publics lie in the range $h \in [0.7, 0.95]$ (Bradshaw & Howard, 2019; Baumann et al., 2020); this is the range swept in the experiments reported in Section 5.

Two network topologies are used. The baseline is a stochastic block model (SBM) with intragroup edge probability p_{in} and intergroup edge probability p_{out} , with $p_{out}/(p_{in} + p_{out}) = 1 - h$ and with a mean degree set to 8 in the experiments. The robustness specification uses a scale-free network generated via the Barabási–Albert preferential attachment procedure, with type assignments randomized over node identifiers. The scale-free specification is chosen because it produces heavy-tailed degree distributions and well-defined high-betweenness bridge edges by construction; this is the topology most favorable to the predicted advantage of targeted adversarial attacks tested in Section 5.

4.2. Agents, State, and Belief Update

Each agent i carries a continuous recognition capacity $\theta_i \in [0, 1]$, drawn at initialization from a $\text{Beta}(\alpha, \beta)$ distribution. The mean of this distribution corresponds to the prior p in the Analytical Model. In the experiments, $\text{Beta}(2, 2)$ is used, which is symmetric and unimodal with mean 0.5, matching the stylized Atomausstieg parameterization (Section 3.5).

Each agent also carries a belief $\mu_i \in [0, 1]$ representing the expected probability that a randomly drawn opposing-group agent is recognition-capable. This summary statistic is the agent-level analog of felt understanding as a Bayesian posterior. Initial beliefs are set to $\mu_i(0) = 0.9$ for all agents; this represents a population starting in the cooperative basin, so that the simulation tests whether the basin is sustainable under perturbation rather than whether cooperation spontaneously emerges from random initial conditions. A robustness check confirms that this choice is consequential rather than innocuous: the system is bistable. From a substantially lower starting belief, $\mu(0) = 0.5$, the population settles into the collapsed basin even at $\pi = 0$, whereas $\mu(0) = 0.9$ sustains cooperation below π^* (Supplementary Figure S2). Fixing $\mu(0) = 0.9$, therefore, isolates the basin sustainability question studied here from the distinct question of whether cooperation spontaneously emerges from a non-cooperative start, which lies outside the present scope.

At each discrete time step t , each agent first observes the signals sent by its cross-group neighbors at $t - 1$, with channel perturbation π applied to each transmission. The receiver's mental model assumes the separating profile of the Analytical Model: type R sends s_A and type N sends s_D . Under this model and the perturbed channel, the log-likelihood ratio of an observed s_A is $\log[(1 - \pi_{eff})/\pi_{eff}]$, where π_{eff} incorporates a noise floor $\varepsilon = 0.05$ to bound the receiver's inference ($\pi_{eff} = \max(\pi, \varepsilon)$). Aggregating across all cross-group neighbor observations in log-odds form and applying a smoothing parameter $\eta = 0.1$ to the update yields the new belief $\mu_i(t)$. Beliefs are bounded away from the extremes (clipped to $[0.05, 0.95]$) to prevent degenerate updating.

Action selection uses a quantal response specification: agent i plays a_R with probability $\sigma(\lambda \cdot ((1 - \pi_{eff}) \cdot u_R \cdot \mu_i - c(\theta_i)))$, where σ is the logistic function, $\lambda = 10$ is the precision parameter, and $c(\theta_i) = c_R + (c_N - c_R)(1 - \theta_i)^2$ is the agent-specific cost interpolating between c_R for $\theta_i = 1$ and c_N for $\theta_i = 0$. As $\lambda \rightarrow \infty$, the quantal response recovers the deterministic best response of the Analytical Model; finite λ introduces realistic stochasticity that prevents absorbing-state lock-in and is standard in behavioral game-theoretic ABMs. The convex (squared) form of the cost penalty encodes that authentic signaling becomes disproportionately costly as recognition capacity falls toward the identity-threatened pole, so that the marginal cost of authentic signaling is small near $\theta = 1$ and steep near $\theta = 0$; a linear interpolation produces the same qualitative dynamics.

4.3. The Operationalization Bridge to Felt Understanding

A methodological clarification is required before the simulation results in Section 5 can be interpreted. The Analytical Model specification defines felt understanding as the Bayesian posterior $\mu_A(\theta_B = R | \hat{s}_B)$ and the simulation tracks this quantity at the agent level as μ_i . However, the posterior has a property that limits its usefulness as the primary outcome measure: in both the separating equilibrium and the pooling equilibrium, the population-averaged posterior equals the prior p . What distinguishes the two regimes is not what agents believe about the other group's underlying type but what agents observably do in cross-group encounters. Below π^* , recognition-capable agents play a_R and emit s_A ; above π^* , all agents pool on a_D and emit s_D . The latent type distribution is identical in both regimes; the action distribution is not.

This distinction maps directly onto the structure of Livingstone et al.'s (2020) empirical scale. The 15 scale items ask respondents to evaluate the observable behavior of the outgroup (in essence, receiver-side judgments about the degree to which outgroup signals convey recognition); for example, the extent to which outgroup members "understand our values," "recognize what we believe in," or "appreciate where we are coming from." These items are not measurements of respondents' inferences about latent psychological types in the outgroup; they are measurements of perceived recognition behavior—the degree to which observable outgroup signals convey acceptance of ingroup perspectives as authentic and valid. In the model, this is the action rate: the fraction of opposing-group agents observed playing a_R , averaged over the agent's cross-group neighbors.

Accordingly, the population-mean action rate is treated as the primary outcome measure and is referred to throughout the Results section as an observable felt understanding proxy. The Bayesian posterior μ_i remains in the model as the mechanism by which agents update beliefs and select actions, but it is the resulting observable behavior—not the underlying belief—that corresponds to what the empirical scale measures. This is the correct methodological alignment: receiver-side recognition behavior is what Livingstone's items operationalize, and it is what the simulation is best positioned to report. Felt understanding as a Bayesian belief about type and felt understanding as observable recognition behavior agree in the no-adversary benchmark but diverge under perturbation; the empirical scale tracks the behavioral quantity, and so do the primary figures reported below. Two clarifications make this bridge explicit. First, in Livingstone et al.'s (2020) terms, felt understanding is a third-order intentional state—an ingroup belief about the outgroup's recognition of ingroup perspectives—and the model does not claim to represent that belief in its full intentional depth; it represents the observable recognition behavior from which the scale's receiver-side judgments are formed. Second, treating the action rate as the proxy is a revealed-preference move of the kind standard in behavioral game theory: because the scale items elicit judgments of what the outgroup observably does, the population action rate is the quantity those judgments track—conditional on the receiver-side calibration as-

sumption that perceived and actual recognition behavior coincide, which we make explicit here and revisit as a limitation in Section 6.4.

4.4. The Adversary

The adversary is parameterized by three values: a perturbation intensity $\pi \in [0, 1/2]$, an injection rate $\gamma \in [0, 1]$ specifying the fraction of cross-group edges into which the adversary injects fabricated dismissive signals each step, and a targeting policy $\tau \in \{\text{uniform, central, bridge}\}$ specifying how the injection budget is distributed across edges.

Uniform targeting distributes injection effort symmetrically across all cross-group edges. Central targeting weights edges by the degree of the receiving node, concentrating perturbation on high-degree agents who are most influential in their local cluster. Bridge targeting weights edges by edge betweenness centrality, concentrating perturbation on edges that span otherwise weakly connected clusters. These three policies test the operational claim, central to the firehose-of-falsehood literature, that strategically targeted disinformation should be more efficient than equivalent-volume uniform disinformation. The total injection budget γ is held constant across policies so that comparisons isolate targeting strategy from raw effort.

4.5. Interventions

Three intervention classes are tested, each corresponding directly to one of the comparative statics levers identified in Section 3.5 of the Analytical Model:

- **Bridge seeding.** A fraction δ of cross-group edges with the highest betweenness centrality are selected, and the agents at both endpoints have their recognition capacity raised to $\theta = 0.9$. This simulates the targeted introduction of recognition-capable advocates at structurally critical positions in the discourse network. The intervention raises the effective prior p in the threshold expression $\pi^* = 1 - c_R/(u_R \cdot p)$, concentrated at structurally consequential locations.
- **Cost reduction.** The cost of authentic signaling c_R is halved for a fraction δ of randomly selected agents. This simulates interventions that lower the social, psychological, or material cost of engaging in cross-group dialogue—for example, facilitated-dialogue platforms, moderated spaces, or media literacy programs. The intervention lowers c_R in the threshold expression.
- **Channel hardening.** The effective perturbation π is reduced by a factor proportional to δ . This simulates interventions that improve the fidelity of cross-group communication channels—platform moderation, authentication systems, or trusted intermediaries. The intervention reduces π directly, attempting to keep the effective perturbation below π^* even under aggressive adversary behavior.

Each intervention is parameterized by a dose $\delta \in [0, 1]$ denoting the fraction of agents or edges affected. The three interventions are tested individually in Section 5; the dose–response of bridge seeding (the most effective) is mapped out in a supplementary experiment.

4.6. Outcome Measures and Validation

For each parameter configuration, the simulation produces three primary outcome measures. The first is the time series of mean action rate within each group, $FU_A(t)$ and $FU_B(t)$, where FU denotes an observable proxy for felt-understanding-supporting behavior as articulated in Section 4.3. The second is the long-run action rate, defined as the mean action rate over the second half of the simulation horizon, averaged across random seeds to control for stochasticity in network construction, type assignment, and action selection. The third is a basin classification: at the simulation horizon, the system is classified as

cooperative if both groups exhibit action rates above 0.3, collapsed if both fall below 0.2, asymmetric if one group collapses while the other does not, and intermediate otherwise.

Before any substantive results are reported, the simulation must pass five validation checks corresponding to the qualitative predictions of the Analytical Model. These are: reduction to the Analytical Model in the appropriate limit (the action rate must drop substantially as π crosses π^*); sustainability of cooperation in the absence of adversary (action rate must remain high when $\pi = 0$); symmetry of outcomes across the two groups under symmetric parameters; monotonic decrease in action rate as π increases through the threshold region; and qualitative robustness of collapse to changes in network topology. All five checks pass; details are reported in the opening of Section 5. The validation suite is treated as a firewall: substantive results from the simulation are reported only after the model has demonstrated faithful reproduction of the Analytical Model's predictions in the appropriate limits.

The simulation is implemented in Python 3.14 using NumPy 2.5.0 for state vectorization, NetworkX 3.61 for graph construction and betweenness centrality computation, and matplotlib for figure generation. The code is reproducible from random seeds and is available with the Supplementary Materials.

5. Results

The results from the Computational Model are reported in four substantive figures plus two supporting analyses. The figures are organized to mirror the manuscript's analytical arc. Figure 2 documents the dynamics of felt-understanding collapse under perturbation. Figure 3 maps the joint dependence of cooperation on adversarial perturbation and network homophily. Figure 4 compares the effectiveness of three policy interventions. Figure 5 examines whether strategic targeting outperforms uniform-volume injection in the adversary's problem, finding that the predicted advantage of strategic targeting is not observed in either topology tested. Figure 6 maps the dose-response curve of the most effective intervention identified in Figure 4. Before presenting the substantive results, the validation outcomes are reported to establish the Computational Model as a faithful extension of the Analytical Model.

5.1. Validation

The five validation checks described in Section 4.6 were executed before any substantive sweeps. The results are summarized below; raw outputs and code are available in the Supplementary Materials. The outcome measure throughout is the population-mean action rate over the second half of the simulation horizon, averaged across random seeds (the observable felt-understanding proxy articulated in Section 4.3).

- Check 1 (the Analytical Model reduction). At the stylized Atomausstieg parameterization ($p = 0.5$, $c_R = 0.35$, $u_R = 1.0$, yielding $\pi^* = 0.30$), the mean action rate at $\pi = 0$ is 0.41 compared to 0.12 at $\pi = \pi^*$, a reduction of 0.29 (five seeds per condition). The simulation reproduces the Analytical Model's qualitative prediction that recognition behavior collapses as the perturbation crosses the analytical threshold.
- Check 2 (no-adversary sustainability). With $\pi = 0$ and no adversarial injection, the mean action rate is 0.44 averaged across five seeds, with per-seed rates ranging from 0.42 to 0.46. Cooperative play is sustained in the absence of perturbation.
- Check 3 (symmetry). Across eight seeds at $\pi = 0.15$, the mean felt-understanding proxy measures for groups A and B are 0.302 and 0.298, respectively, with the difference well within two pooled standard errors. The simulation does not exhibit asymmetries between the two domestic groups under symmetric parameterization.

- Check 4 (monotonicity through the threshold). Across $\pi \in \{0.0, 0.1, 0.2, 0.3\}$, action rate decreases monotonically: 0.37, 0.19, 0.14, and 0.11. The total decline across the analytically relevant range is 0.26.
- Check 5 (specification robustness). At $\pi = 0.35$, both stochastic block model networks (mean action rate 0.12) and Barabási–Albert scale-free networks (mean action rate 0.03) exhibit collapse. The qualitative result does not depend on the choice of network topology.

All five checks pass. The substantive results follow.

5.2. Dynamics of Recognition Collapse

Figure 2 shows the time evolution of the observable felt-understanding proxy for both domestic groups at three perturbation regimes: well below the analytical threshold ($\pi = 0.05$), approaching the threshold ($\pi = 0.15$), and at the threshold ($\pi = 0.30$). Each trajectory averages three random seeds and is smoothed with a centered rolling mean of width 10.

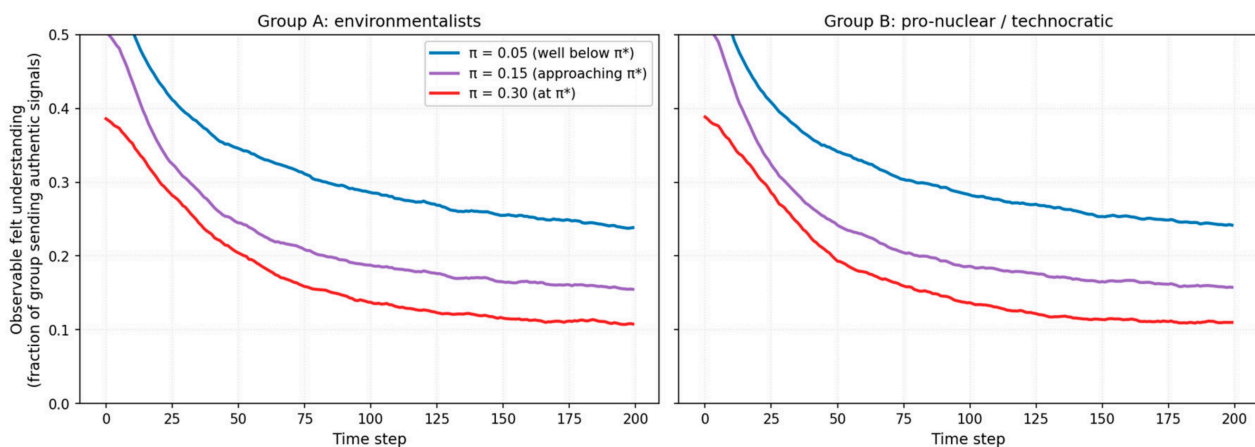


Figure 2. Cross-group recognition behavior under three perturbation regimes. Each curve shows the smoothed mean action rate within the indicated group, averaged across three random seeds. Higher π produces lower sustained recognition behavior; the ordering is monotone and consistent across both groups. The system equilibrates from the high-initial-belief condition within ≈ 100 time steps; the steady-state ordering reflects the Analytical Model’s comparative statics prediction. The horizontal axis is simulation time (steps) and the vertical axis is the observable action rate (the fraction of the indicated group sending authentic, recognition-conveying signals); the left and right panels correspond to groups A and B. The three solid curves are perturbation regimes below, approaching, and at the analytical threshold: $\pi = 0.05$ (blue), $\pi = 0.15$ (purple), and $\pi = 0.30 = \pi^*$ (red).

Three features of these trajectories are worth noting. First, all three regimes show an initial transient as the system equilibrates from the high-belief initial condition; this transient completes within roughly 100 time steps. Second, the steady-state ordering is exactly as the Analytical Model predicts: lower π sustains higher cooperative behavior, and the two domestic groups remain symmetrically positioned throughout. Third, the gap between the below-threshold regime ($\pi = 0.05$, sustaining ≈ 25 percent cooperative play) and the at-threshold regime ($\pi = 0.30$, sustaining ≈ 11 percent) represents a roughly 60 percent reduction in observable felt understanding from a moderate adversarial perturbation. Even before the analytical threshold is decisively crossed, signal degradation substantially erodes cross-group recognition behavior.

5.3. The Phase Geometry of Collapse

Figure 3 maps the long-run action rate across the joint parameter space of adversarial perturbation π and network homophily h . The grid spans $\pi \in [0, 0.45]$ and $h \in [0.5, 0.95]$ (seven and six grid points, respectively), with two random seeds per cell. The analytical threshold $\pi^* = 0.30$ from the Analytical Model is marked.

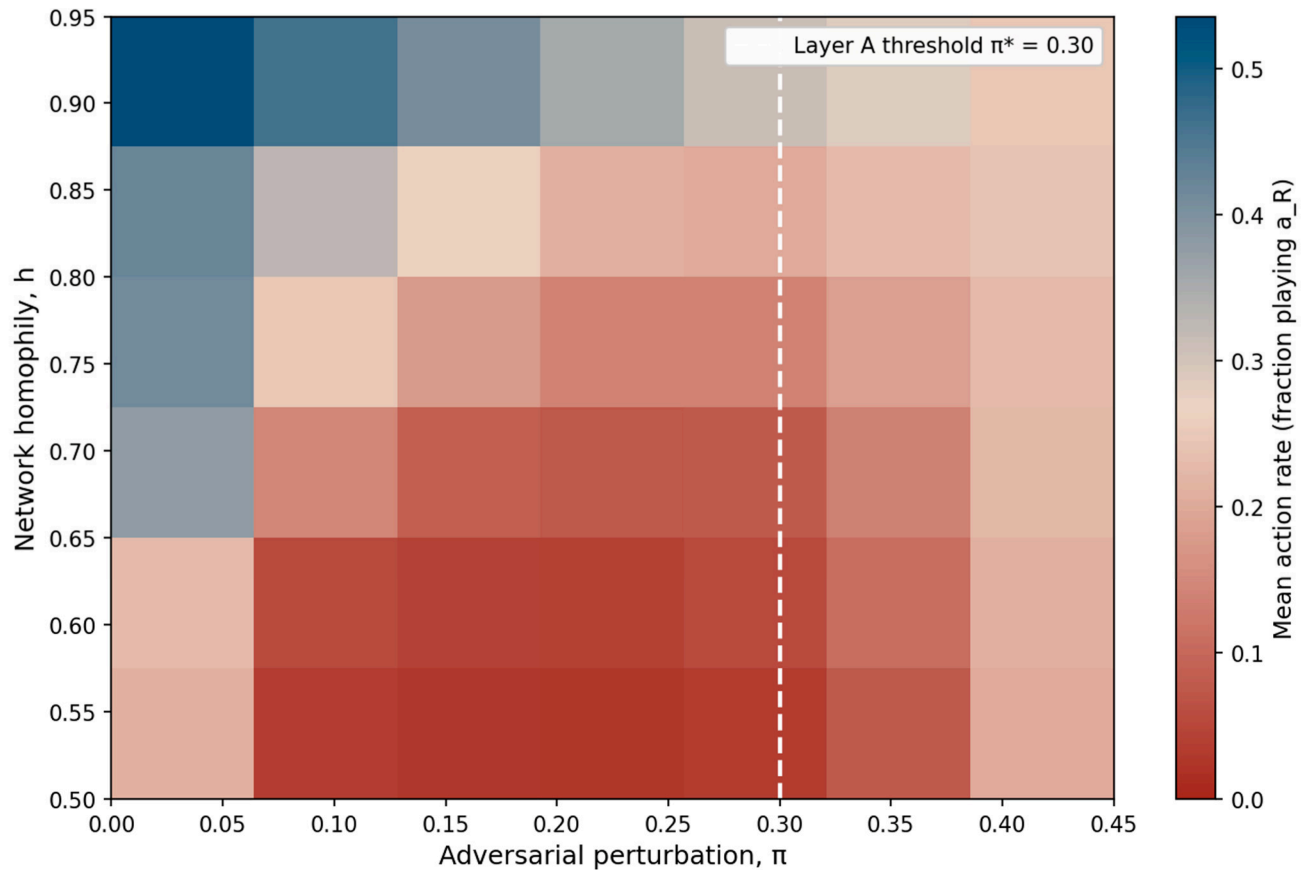


Figure 3. Long-run action rate across the (π, h) parameter space. Blue cells indicate parameter combinations sustaining recognition behavior; red cells indicate collapse. The vertical white dashed line marks the analytical threshold $\pi^* = 0.30$ from the Analytical Model. High homophily ($h \geq 0.85$) protects cooperation below the threshold but provides no rescue above it; once π crosses the threshold, no parameter combination tested produces a cooperative basin. The horizontal axis is the adversarial perturbation π , and the vertical axis is network homophily h . The color scale is continuous and encodes the long-run action-rate magnitude: saturated blue denotes sustained recognition, saturated red denotes collapse, and intermediate shades denote partial cooperation.

Three findings emerge from the phase diagram analysis (Figure 3). First, the analytical threshold π^* from the Analytical Model identifies the perturbation level at which long-run cooperation begins to break down across all homophily levels tested. This is the primary validation of the formal model: a static analytical result correctly predicts the boundary of collapse in the dynamic networked simulation.

Second, network homophily has a substantial protective effect at low perturbation. Highly homophilic networks ($h \geq 0.85$) sustain cooperative behavior at perturbation levels where weakly homophilic networks ($h \leq 0.65$) have already collapsed. The mechanism is transparent in the model: within-group clustering allows recognition-capable agents to reinforce each other's cooperative play before that play is diluted by cross-group exposure to dismissive signals. This finding is counterintuitive in the broader polarization literature, which typically treats homophily as universally bad for democratic dialogue. The present

result qualifies that conclusion: homophily is a substrate that stabilizes whatever equilibrium the system is currently in. Below the threshold, stabilization protects cooperation; above the threshold, it accelerates entrenchment in the collapsed basin.

Third, no parameter combination tested produces a fully cooperative basin at high π . The upper portion of Figure 3 is uniformly red across the homophily range, confirming the Analytical Model prediction that once the channel is sufficiently corrupted, no network structure rescues the cooperative regime. The implication for the Atomausstieg case is direct. Pre-2011 German discourse around energy policy was characterized by high homophily within the environmentalist and pro-nuclear publics but with functioning cross-group dialogue (Fairchild, 2026). The model suggests that this combination was sustainable up to a perturbation threshold but could not survive once disinformation infrastructure pushed effective π past that threshold. The homophily that protected the cooperative basin during the pre-perturbation period became, once the threshold was crossed, an accelerator of within-group radicalization with no cross-group corrective.

5.4. Intervention Effectiveness

Figure 4 compares the three intervention strategies derived from the comparative statics of Proposition 1's threshold expression $\pi^* = 1 - c_R/(u_R \cdot p)$. Each intervention is applied at a 20 percent dose—affecting 20 percent of agents or edges depending on the intervention type—across the perturbation range $\pi \in [0.05, 0.40]$.

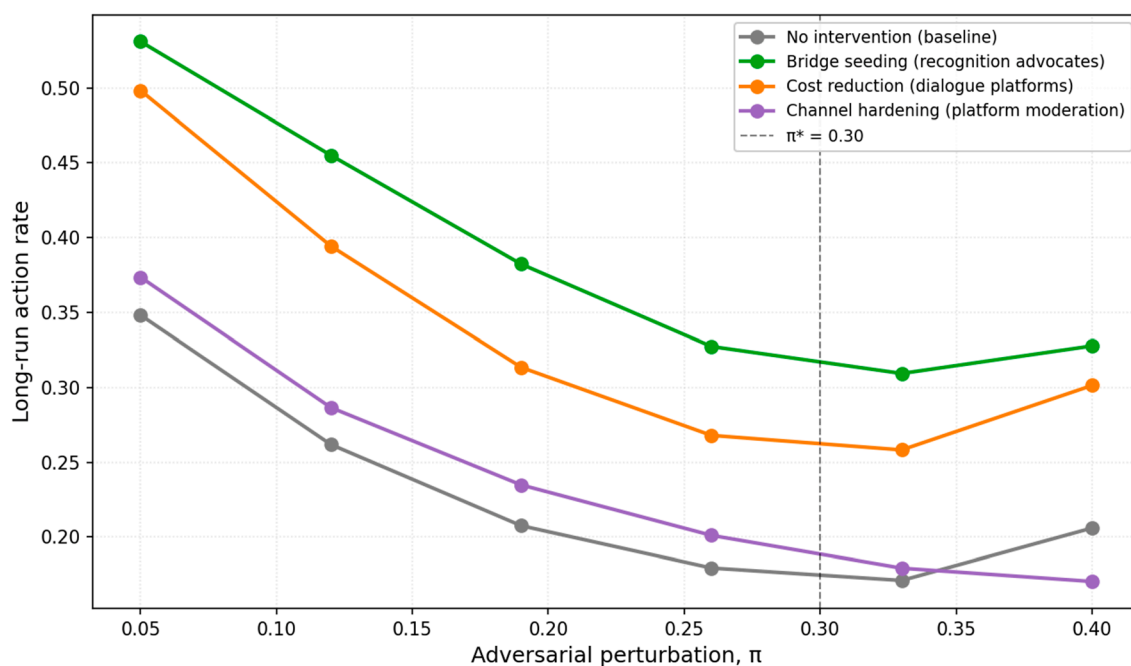


Figure 4. Intervention effectiveness across the perturbation range. Bridge seeding (placement of recognition-capable agents at structurally central cross-group edges) is the most effective intervention at every perturbation level tested. Cost reduction ranks second; channel hardening provides only modest improvement over the baseline. The vertical dashed line marks $\pi^* = 0.30$. The horizontal axis is the adversarial perturbation π , and the vertical axis is the long-run action rate; each solid colored line with markers traces one intervention class, while the single vertical dashed line marks the analytical threshold π^* .

The effectiveness ordering is stable across the perturbation range and substantively striking: bridge seeding outperforms cost reduction, which outperforms channel hardening, which provides only modest gains over the no-intervention baseline. At the threshold $\pi = 0.30$, bridge seeding sustains an action rate of ≈ 0.31 compared to the baseline rate of

0.17 at the same perturbation level—nearly a doubling of observable felt understanding from a 20 percent intervention dose. Cost reduction recovers to ≈ 0.26 at the same point. Channel hardening, the intervention class corresponding to platform moderation, lifts the action rate only modestly above baseline.

The intuition is structural. Bridge seeding directly raises the effective prior p , and it does so at locations where the seeded agents' cooperative play propagates through the network most efficiently. Cost reduction lowers c_R but distributes the effect uniformly across the population, including across agents whose network positions are not consequential for cross-group transmission. Channel hardening lowers effective π but does so against an adversary whose injection budget is high enough that partial hardening leaves the effective perturbation well above π^* . The threshold comparative statics decomposition of Proposition 1 thus generates the ordering: interventions that raise the right side of the inequality $\pi < 1 - c_R/(u_R \cdot p)$ by changing p (bridge seeding) outperform those that change c_R (cost reduction), which outperform those that attempt to reduce the left side directly (channel hardening).

This result has direct policy implications, taken up in Section 6. The dominant policy response to disinformation in Western democracies has been platform moderation—the channel-hardening intervention. The model identifies it as the least effective of the three classes tested.

The effectiveness ordering is robust to the stylized parameterization. Re-running the intervention comparison across six combinations of the prior, the authentic-signaling cost, and the recognition payoff—spanning prior values $p \in \{0.40, 0.50, 0.60\}$, cost values from 0.25 to 0.45, and payoff up to 1.3, each evaluated at its own analytical threshold π^* —shows that the intervention ranking is substantively preserved. Bridge seeding is the strongest intervention in nearly every condition and is numerically tied with cost reduction at the single high-prior point ($p = 0.60$), where cost reduction exceeds bridge seeding only trivially in the mean. Channel hardening remains the weakest active intervention across the tested combinations and, in one condition, falls below the no-intervention baseline. The ordering, therefore, follows from the comparative statics structure of Proposition 1 rather than from the specific numerical values and does not depend materially on the stylized parameter choices (Supplementary Figure S1).

5.5. Adversary Targeting: Volume Versus Precision

Figure 5a examines the strategic question on the adversary side. The model is held at $\pi = 0.15$ (below the threshold), and the adversary's injection rate γ is varied across $[0.0, 0.4]$ under three targeting policies: uniform, central (weighted by receiver degree), and bridge (weighted by edge betweenness centrality). The substantive prediction underlying the firehose-of-falsehood literature is that strategically targeted disinformation should be more efficient at degrading cross-group dialogue than equivalent-volume uniform disinformation. To isolate targeting strategy from raw effort, the total injection budget γ is held exactly equal across policies by sampling, at each step, exactly $\text{round}(\gamma \cdot M)$ directed cross-group transmissions without replacement, with sampling probabilities proportional to the targeting weights.

On the dense block-structured network, the three targeting policies do not differentiate sharply. Uniform targeting is most damaging at most injection rates, but the gaps across policies (typically 0.003 to 0.015 in action rate at $\gamma \geq 0.16$, with standard errors of ≈ 0.004 to 0.007) are at most one to two standard errors. Bridge targeting is consistently slightly less damaging than uniform across the mid-to-high γ range; central targeting is intermediate. The qualitative finding is that strategic precision provides no systematic advantage over equivalent-volume uniform injection on the SBM, and in fact, bridge targeting underper-

forms uniform by a small but persistent margin. The mechanism is structural. Within a tightly clustered homophilic group, the average path length between any two nodes is short, and any single agent's exposure to dismissive signals diffuses through the local neighborhood regardless of where the adversary concentrates the injection. When the injection budget is held constant, concentrating that budget on high-betweenness edges does not produce more cross-cluster damage than spreading it across all cross-group transmissions systematically; if anything, the precision approach wastes injection mass on edges already saturated by their local context.

A reasonable concern with this finding is that it may be specific to the dense block-structured topology. Scale-free networks are the natural counterexample because they exhibit heavy-tailed degree distributions and well-defined high-betweenness bridge edges by construction—the topological structure that centrality-based targeting is designed to exploit. Figure 5b reruns the targeting comparison on a Barabási–Albert scale-free network with $N = 5000$ to test whether the SBM finding generalizes.

The scale-free result is consistent with and slightly sharper than the SBM result. At moderate injection rates ($\gamma = 0.10$ to 0.20), uniform targeting is detectably more damaging than central or bridge targeting, with the gap reaching approximately two standard errors. At higher injection rates, the policies converge as the injection budget saturates the channel. Across the full injection range, central and bridge targeting are statistically indistinguishable from one another. The predicted strong advantage of strategically targeted disinformation is not observed; if anything, uniform targeting is marginally more damaging than precision targeting at the perturbation rates where policy differentiation is detectable. This finding qualifies the dominant operational claim underlying the firehose-of-falsehood literature in an interesting way: strategic targeting is not catastrophically worse than uniform injection (a sophisticated adversary employing precision tactics would still succeed in degrading cross-group recognition), but it is also not better than equivalent-volume uniform injection. The intuition is that under realistic homophily, the structural advantage that high-betweenness positions confer on the receiver side (rapid diffusion of received signals across local neighborhoods) is approximately offset by the inefficiency of concentrating injection effort on those same positions on the sender side, when an equivalent budget could have been spread across the entire cross-group transmission surface.

The direction of this result—uniform injection weakly dominating targeted injection under an equal budget—can be understood as a concavity effect in the adversary's per-edge damage function; Appendix B states the corresponding constrained optimization problem. Let $d(m)$ denote the contribution to long-run misrecognition from directing injection mass m at a given cross-group edge. Two features of the model's specification make d increasing but concave in m . First, receivers update beliefs in log-odds space with bounded support: each observed dismissive signal shifts the receiver's log-odds by a fixed increment that is then smoothed (η) and clipped to $[0.05, 0.95]$, so the marginal effect of an additional dismissive signal on a receiver's belief—and hence on its cooperative-action probability—diminishes as that receiver is driven toward the dismissive pole and vanishes once the clip binds. Second, the quantal-response action rule $\sigma(\cdot)$ is itself sigmoidal, so for a receiver already in the dismissive regime, further suppression of its belief yields a vanishing additional reduction in an already-low cooperation probability. The composition of a saturating belief update with a saturating action rule makes d concave in m . With the total injection mass across cross-group edges held fixed by the equal budget constraint, and per-edge damage functions approximately common across edges, concavity implies—by Jensen's inequality—that the aggregate damage, the sum of $d(m)$ over all cross-group edges, is maximized by spreading mass as evenly as possible and is reduced by any concentration of mass onto a subset of edges. Targeting (central or bridge) is exactly such a

concentration; it overcomes the concavity penalty only if the targeted high-betweenness edges carry a sufficiently higher damage ceiling or steeper initial slope. Under realistic homophily, they do not: short intra-cluster path lengths mean that a dismissive signal entering a cluster through any edge diffuses to the whole cluster within a few steps, so the receiver-side diffusion premium of entering through the single highest-betweenness edge over a typical edge is small, while the per-edge damage ceiling is capped by the same belief clipping that generates the concavity. The targeting premium is thus second-order while the concentration penalty is first-order, and uniform injection wins by the resulting small margin. The same logic explains the topology dependence of the effect's magnitude: on the scale-free network, a few high-degree hubs absorb concentrated injection and reach the clip quickly, so targeting wastes mass on saturated hubs at moderate injection rates—which is why uniform's advantage is detectable ($\approx$ two standard errors) at $\gamma = 0.10$ – 0.20 yet disappears at $\gamma \geq 0.30$, where the budget is large enough that even uniform injection saturates the channel and the allocation ceases to matter. The result is, therefore, not an artifact of a single topology; it is a generic consequence of saturating damage under a fixed budget, with the dense homophilic SBM and scale-free cases differing primarily in how quickly saturation is reached.

Two implications follow. First, the empirical preference of Russian state-sponsored information operations for high-volume, low-precision tactics over surgical targeting (Paul & Matthews, 2016) is consistent with the structure of the informational environment in which such operations are deployed: an adversary indifferent between uniform and targeted approaches, on the evidence of these simulations, will rationally prefer the approach that requires less intelligence and operational overhead. Second, defensive responses to disinformation that focus on identifying and protecting bridge nodes specifically (rather than on raising the structural cooperation threshold more broadly) may not be addressing the dominant attack mode. These conclusions hold conditionally on the topologies and parameter ranges tested; whether they generalize to other network classes (small-world, multiplex, evolving) is a question for further work.

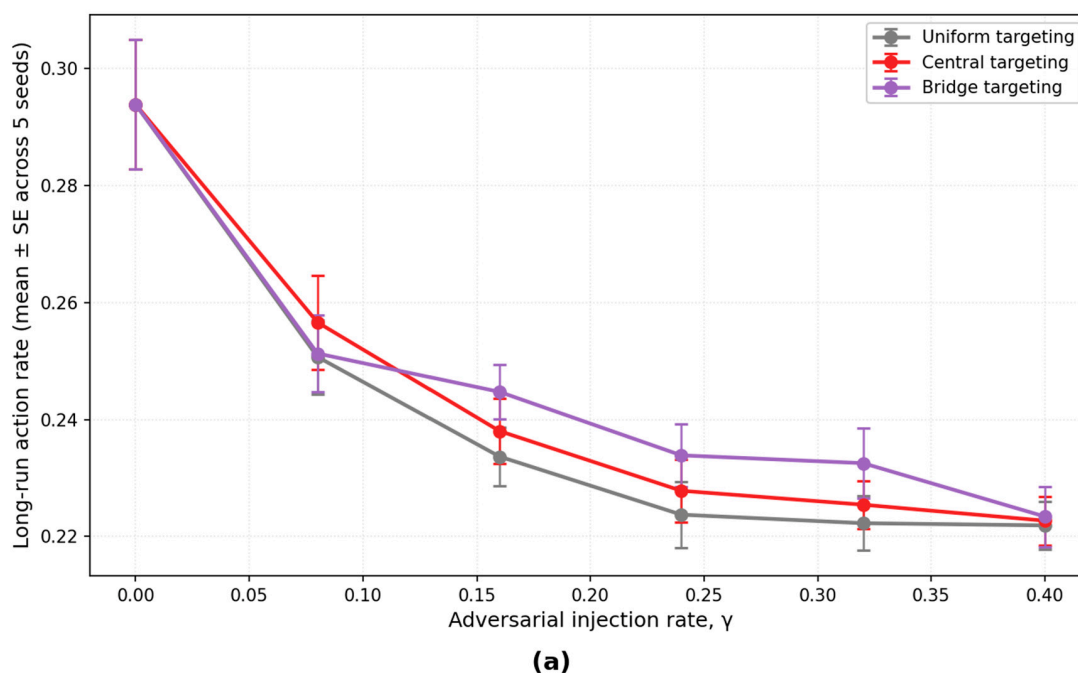


Figure 5. Cont.

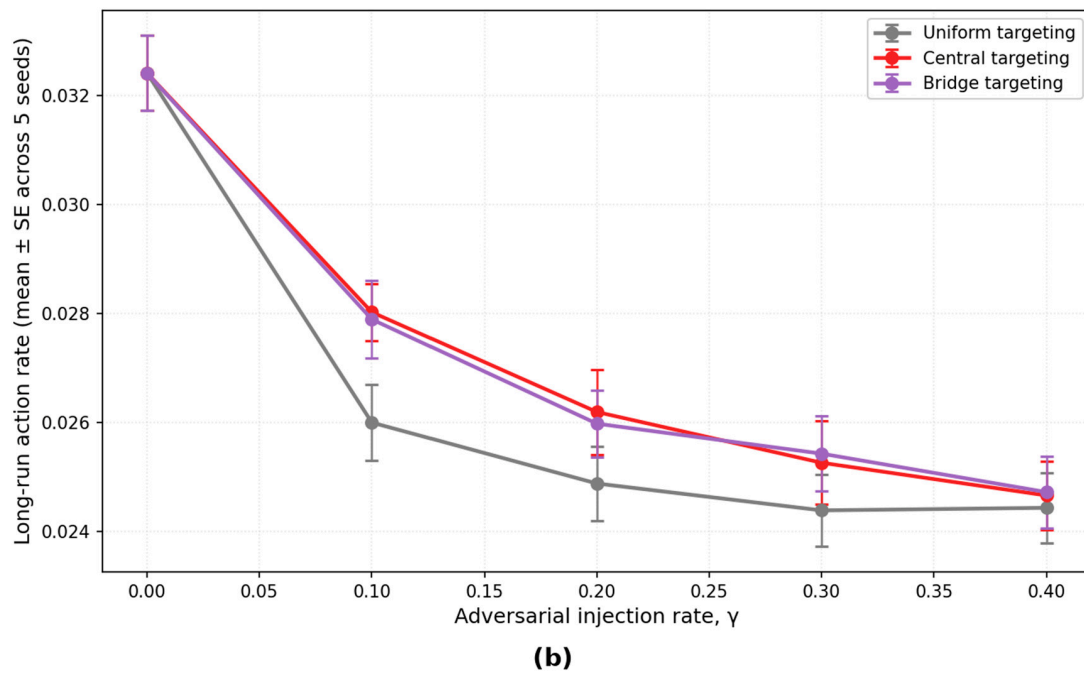


Figure 5. (a) Adversary targeting policies under an exact-count injection budget. In both panels, the baseline perturbation is fixed at $\pi = 0.15$, the injection rate γ is varied across the three targeting policies, and error bars show ± 1 standard error across five random seeds per cell. (a) Small stochastic block model network ($N = 300$, mean degree 6). The three policies produce statistically similar levels of damage across the injection range, with uniform targeting marginally most damaging at most injection rates, but central and bridge targeting overlapping uniformly within one to two standard errors. The predicted strong advantage of strategically targeted disinformation is not observed. (b) Large scale-free network ($N = 5000$, Barabási–Albert generator, mean degree 8) with type assignments randomly permuted relative to node identifiers to prevent degree–type correlation. At moderate injection rates ($\gamma = 0.10, 0.20$), uniform targeting is detectably more damaging than central or bridge targeting (a gap of ≈ 0.001 to 0.002 in action rate, roughly two standard errors). At high injection rates ($\gamma \geq 0.30$), the three policies converge. Central and bridge targeting are statistically indistinguishable from one another throughout. Baseline action rates are low (≈ 0.024 to 0.032), reflecting the difficulty of sustaining cross-group cooperation in a large heterogeneous scale-free topology even at low perturbation.

5.6. Bridge Seeding Dose–Response

Figure 4 establishes that bridge seeding is the most effective of the three intervention classes at a fixed 20 percent dose. Figure 6 maps the dose–response curve of bridge seeding across the dose range $\delta \in [0, 0.4]$ at three perturbation levels: below the threshold ($\pi = 0.15$), approaching the threshold ($\pi = 0.25$), and above the threshold ($\pi = 0.35$).

The dose–response is approximately linear across all three perturbation regimes, with no apparent saturation in the range tested. The intervention is effective even above the analytical threshold, where Proposition 1 predicts that the unaided system is in the pooling basin. At $\pi = 0.35$, the baseline action rate is 0.18; a 30 percent dose of bridge seeding lifts the action rate to ≈ 0.38 , more than doubling observable felt understanding.

This result extends the policy claim from Section 5.4. Bridge seeding is not merely the most effective intervention at below-threshold perturbation levels; it can actively rescue cooperation when the system has been pushed into the pooling basin by adversarial activity. The mechanism is the same in both regimes—targeted introduction of recognition-capable agents at structurally critical positions raises the effective prior p —but the substantive implication is stronger above the threshold: intervention can do what the unaided system

cannot, namely, escape the basin of misrecognition that Proposition 1 identifies as the unique equilibrium under high perturbation.

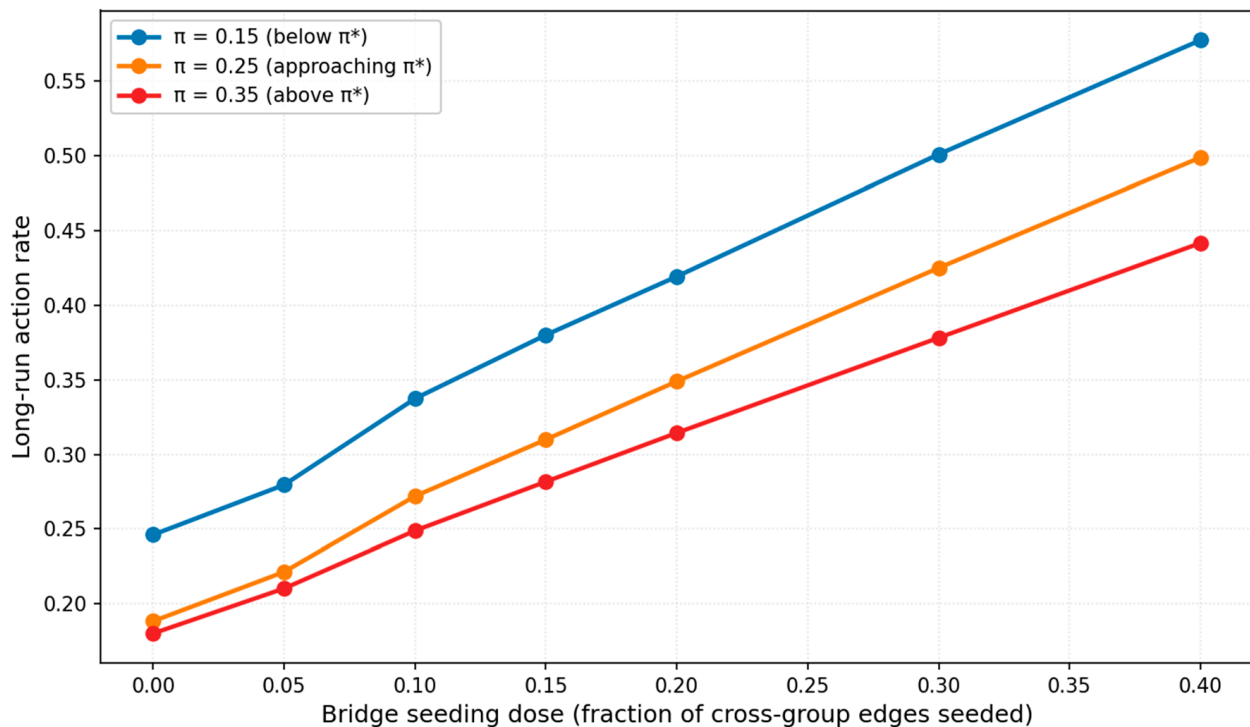


Figure 6. Bridge seeding dose-response across perturbation regimes. The response is approximately linear in dose across all three regimes. Crucially, bridge seeding rescues cooperation even above the analytical threshold: at $\pi = 0.35$ (above $\pi^* = 0.30$), a 30 percent dose lifts the action rate from a baseline of 0.18 to ≈ 0.38 , more than doubling observable felt understanding. The horizontal axis is the bridge seeding dose (the fraction of cross-group edges seeded), and the vertical axis is the long-run action rate; each solid line corresponds to one perturbation regime, with lower π regimes lying higher.

6. Discussion

Four findings emerge from the model. First, the analytical threshold $\pi^* = 1 - c_R / (u_R \cdot p)$ derived in Proposition 1 organizes the simulation's behavior across the full parameter space tested in Section 5. The static signaling game and the dynamic agent-based simulation agree on the boundary of collapse; the threshold is the substantive predictor of long-run cooperative behavior. Second, network homophily protects cooperation below the threshold but offers no rescue above it; the topological feature that stabilizes the cooperative basin also stabilizes the collapsed basin once the threshold is crossed. Third, of three intervention classes tested—bridge seeding, cost reduction, and channel hardening—bridge seeding is consistently most effective, and its dose-response is linear enough that it rescues cooperation even above π^* . Fourth, and contrary to a substantive prediction of the firehose-of-falsehood literature, uniform adversarial volume and strategically targeted adversarial precision produce approximately equivalent damage across both dense block-structured and large scale-free network topologies, with uniform marginally more damaging at most injection rates, but the gap is small relative to between-seed variance. The model suggests that the observed operational preference of state-sponsored disinformation for high-volume low-precision tactics is consistent with this approximate equivalence: an adversary indifferent between approaches will rationally prefer the one requiring less intelligence overhead.

These findings bear on the Atomausstieg case, on psychological theories of intergroup recognition, on policy responses to disinformation, and on the boundaries of the present model. Each is addressed in turn.

6.1. Implications for the Atomausstieg Case

The German Atomausstieg fits the model's phase diagram in Figure 3 with uncomfortable precision. Pre-2011 German energy discourse exhibited the combination of high within-group homophily and functioning cross-group dialogue that locates the system in the blue cooperative basin at low π (Fairchild, 2026). The buildup of Russian state-sponsored disinformation infrastructure through Gazprom-linked NGOs, aligned media properties, and platform-amplified messaging across the 2011–2022 window (Bennhold & Solomon, 2022; Harvey, 2014) corresponds in the model to increasing π . The 2011 Fukushima accident functioned as a narrative accelerant that amplified the dread risk frame exploited by the disinformation apparatus (Fairchild, 2026; Knight, 2011; Stamp, 2011), shifting the operational environment toward the threshold faster than the disinformation effort alone would have done.

The model predicts the observed downstream outcome. Once π crossed π^* in the model, the cooperative basin between the environmentalist and pro-nuclear publics became unsustainable. The same network homophily that had previously stabilized in-group cooperation under conditions of light perturbation became, under heavy perturbation, the substrate for rapid within-group radicalization with no functioning cross-group corrective. The empirical record matches: the Energiewende failed to decarbonize, lignite generation increased, fossil gas dependence on Russia deepened, an estimated 800 excess deaths per year resulted from the air pollution that displaced nuclear generation (Jarvis et al., 2022), and Germany's initial response to the 2022 invasion of Ukraine was constrained by an energy posture that the model identifies as an equilibrium outcome of a decade of engineered misunderstanding.

The model does not claim that disinformation was the sole cause of the Atomausstieg trajectory. Domestic political dynamics, industrial policy preferences, and the genuine intensity of the post-Fukushima anti-nuclear sentiment all played independent roles. What the model suggests is that the cross-group felt-understanding collapse documented in the Fairchild (2026) case study is consistent with a threshold-crossing event of the kind Proposition 1 characterizes, and that disinformation operations of the volume and operational style empirically documented for the period are sufficient to push the effective π across that threshold.

6.2. Clinical Psychology Implications: Recognition Advocacy

The simulation's strongest intervention result is that bridge seeding—the targeted placement of recognition-capable agents at structurally critical cross-group network positions—outperforms both cost reduction (dialogue platforms) and channel hardening (platform moderation) at every perturbation level tested, and it rescues cooperation even above the analytical threshold (Figures 4 and 6). The formal mechanism is transparent: bridge seeding raises the effective prior p in the threshold expression $\pi^* = 1 - c_R / (u_R \cdot p)$, concentrated at network positions where the raised effective p propagates most efficiently. The substantive interpretation is more interesting.

The bridge seeding intervention class has a conceptual analog in clinical and humanistic psychological theory on recognition. Rogers (1995) developed the construct of unconditional positive regard as a relational stance in which one party communicates acceptance of the other's subjective perspective without requiring agreement. Reis et al. (2017) situate feeling understood within the interpersonal recognition literature as a core relational

good. Livingstone et al.'s (2020) operationalization of felt understanding builds on this lineage by refining interpersonal feeling-understood research for intergroup contexts and distinguishing felt understanding from feeling liked or agreed with. The bridge seeding intervention in our model corresponds operationally to the targeted introduction of agents who embody the recognition stance developed in this clinical tradition into structurally consequential positions in the cross-group discourse network.

Two implications follow. First, recognition advocacy at scale is not optional for democracies seeking to defend against engineered misunderstanding; it is the highest-leverage intervention class the model identifies. Second, the placement of recognition advocates matters as much as their existence. The model's bridge seeding outperforms uniform cost reduction precisely because seeding is targeted: high-recognition agents placed at high-betweenness positions in the cross-group network produce substantially more cooperative recovery than the same number of high-recognition agents placed randomly. This suggests that the design of recognition-restoring interventions should attend not only to the training and cultivation of recognition-capable individuals but also to the structural positioning of those individuals within the discourse network they are meant to stabilize.

6.3. Policy Implications: The Limits of Platform Moderation

The dominant Western policy response to state-sponsored disinformation has been some variant of platform moderation: content-policy enforcement, deplatforming of inauthentic accounts, labeling of state-affiliated media, and algorithmic deamplification of identified disinformation sources. In the present model, this intervention class corresponds to channel hardening: reducing the effective π on a fraction of cross-group communication channels. The model identifies it as the least effective of the three intervention classes tested. A complementary, governance-side perspective is offered by Lei et al. (2026), whose evolutionary game model of platform–user–government interaction characterizes the conditions under which moderation, user participation, and regulation jointly converge on effective disinformation governance—a layer of strategic response that operates above the recognition dynamics modeled here.

The mechanism of the underperformance is structural. Channel hardening attempts to reduce the left side of the threshold inequality $\pi < 1 - c_R/(u_R \cdot p)$ directly. But the adversary's injection budget is high—this is the operational signature of the firehose model—and partial channel hardening leaves the effective perturbation well above π^* across most of the parameter range tested. Worse, channel hardening is an attritional intervention: each blocked account, labeled article, or deamplified link consumes operational resources for a counterparty whose marginal cost of creating new channels of injection is effectively zero. The targeting results from Section 5.5 compound the difficulty in a different way: because volume-based and precision-based injection produce approximately equivalent damage on the topologies tested, an adversary facing channel-hardening pressure can substitute one operational mode for the other without losing effectiveness, while a defender investing in channel-hardening capabilities specific to one mode gains no leverage against the other.

The model does not argue that platform moderation is useless. It argues that platform moderation is the wrong primary intervention and that resources allocated to it should be considered in light of the higher-leverage alternatives the comparative statics identify. Bridge seeding raises p ; cost reduction lowers c_R ; both attack the right side of the threshold inequality, where the adversary cannot directly counter the intervention. A policy response to disinformation calibrated to the structure of the model would invest substantially more in the cultivation and structural placement of recognition advocates, somewhat less in dialogue infrastructure and media literacy, and substantially less in platform moderation than the current Western policy mix implies.

6.4. Limitations

Five limitations of the present analysis should be mentioned. First, the adversary is non-strategic. The adversary is modeled as a parameter of the environment that determines π and the targeting policy, rather than as a third strategic player with its own optimization problem. Section 3.6 justifies this choice as the standard Stackelberg leader reduction and shows that the present comparative statics already characterize the adversary's objective surface, formalized in Appendix B. A natural extension would model the adversary as choosing π and τ to maximize expected cross-group misrecognition subject to a resource constraint, producing a three-player game whose equilibrium analysis is more complex but more faithful to the strategic logic of state-sponsored disinformation operations. This extension should be anticipated in subsequent research.

Second, the binary type space in the Analytical Model is a simplification of the continuous recognition capacity that the empirical scale measures. The Computational Model addresses this with continuous $\theta \in [0, 1]$, and the validation suite verifies that the Computational Model reduces to the Analytical Model in the binary limit, but the closed-form threshold expression of Proposition 1 is specific to the binary specification. A continuous-type analytical model is a worthwhile extension.

Third, the bridge between receiver-side empirical measurement and sender-side formal model is interpretive rather than direct. Livingstone et al.'s (2020) scale measures receiver judgments; the present model is about sender signaling and the channel that mediates between them. The bridge depends on the assumption that receiver judgments are well-calibrated to the observable behavior of the outgroup, which the operationalization in Section 4.3 makes explicit but which is not directly tested by the empirical literature.

Fourth, the model treats the discourse environment as closed. There is no exogenous shock—a Fukushima-scale event, an election cycle, a security crisis—that could shift the prior or the cost structure independently of signaling dynamics. The Fairchild (2026) case study emphasizes that Fukushima functioned as a narrative accelerant interacting with pre-existing dread risk; modeling such accelerants would require an exogenous shock process layered onto the present model. This is a natural extension for follow-up work.

Fifth, within-group heterogeneity in the present model is limited to differences in recognition capacity θ and network position. Real disinformation operations exploit additional heterogeneity in demographics, ideology, platform usage, and prior exposure. Extending the model to richer agent heterogeneity is straightforward but expands the parameter space substantially and is set aside for the baseline analysis.

7. Conclusions

This analysis proposes a two-layered formal model of engineered misunderstanding. The Analytical Model, a static Bayesian signaling game with a noisy adversarial channel, establishes a closed-form threshold $\pi^* = 1 - c_R / (u_R \cdot p)$ above which the unique Perfect Bayesian Equilibrium is mutual misrecognition. The Computational Model, an agent-based extension of homophilic and scale-free networks, validates the Analytical Model's threshold in the appropriate limit and extends the analysis to dynamics, network structure, and intervention effectiveness. Together, the two layers identify four substantive results: the analytical threshold organizes the phase space of cooperative outcomes; high network homophily protects cooperation below the threshold but accelerates collapse above it; bridge seeding outperforms cost reduction and channel hardening as an intervention class and rescues cooperation even above the threshold; and uniform adversarial volume and strategically targeted adversarial precision produce approximately equivalent damage across both dense and scale-free topologies, qualifying rather than confirming the strong form of the firehose-of-falsehood doctrine while remaining consistent with the observed

operational preference of state-sponsored disinformation for high-volume low-precision tactics.

The model is anchored narratively to the German Atomausstieg case documented in Fairchild (2026) and identifies an equilibrium mechanism consistent with the cross-group recognition collapse documented in that case. The psychological implications indicate the importance of recognition advocacy in the lineage of Rogers (1995) and Reis et al. (2017) as the highest-leverage policy intervention class tested in the model. The policy implications cut against the prevailing Western emphasis on platform moderation as the primary response to disinformation.

Three directions for follow-up work are signaled by the limitations discussion. A strategic adversary with its own optimization problem would convert the two-player formal model into a three-player game whose equilibrium analysis would more faithfully capture the operational logic of state-sponsored disinformation. A continuous-type closed-form extension would generalize Proposition 1's threshold to the continuous recognition capacity that the empirical scale measures. An exogenous-shock extension would model narrative accelerants of the Fukushima type explicitly, capturing the interaction between dread-amplifying events and adversarial perturbation that the case study identifies as central. Each represents an opportunity to extend the analysis without abandoning the structural insights established here.

The broader implication is that the collapse of cross-group felt understanding under engineered misunderstanding is not a transient confusion to be cleared up by better fact-checking or more aggressive platform enforcement. It is an equilibrium outcome of a strategic environment in which an adversary holds the operational capacity to push the perturbation parameter past the point at which the cooperative basin can no longer hold. Recognizing this collapse as an equilibrium rather than an error reframes what counts as a serious response: the task is not to correct false claims after they land but to preserve the structural conditions under which cross-group recognition remains possible at all. Fifth-generation warfare should not be understood merely as an informational or policy problem—it is a population-level psychological stressor implicated in depression, trauma exposure, and symptoms of post-traumatic stress disorder (Jobson et al., 2023; Krishnan, 2024), and those vulnerabilities sharpen precisely when polarized groups lose the felt sense that their perspectives are recognized as authentic or subjectively valid by the other side (Livingstone et al., 2020). The model identifies the levers that push the threshold outward; what remains is the political work of pulling them, before engineered misunderstanding stabilizes into the new equilibrium it was designed to produce.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/g17040036/s1>. Figure S1: Sensitivity of intervention effectiveness to the stylized parameterization. Long-run action rate under bridge seeding, cost reduction, channel hardening, and the no-intervention baseline, recomputed across six combinations of the prior p , the recognition cost cR , and the recognition payoff uR . The intervention ranking is substantively preserved across the tested combinations: bridge seeding is the strongest intervention or numerically tied with cost reduction, while channel hardening remains the weakest active intervention. Figure S2: robustness of the collapse dynamics to the initial belief $\mu(0)$ —long-run action rate as a function of $\mu(0)$ at $\pi = 0$, showing that the system is bistable. Trajectories initialized near the cooperative basin ($\mu(0) = 0.9$) remain cooperative, whereas trajectories initialized at the uninformative prior ($\mu(0) = 0.5$) settle into the collapsed basin.

Funding: This research received no external funding.

Data Availability Statement: The code and data supporting the findings of this study, including the agent-based model and the scripts that reproduce every figure, are openly available at

https://github.com/rrlresearch/Fairchild_Engineered_Misunderstanding_GamesMDPI (accessed on 25 June 2026).

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. Complete Proof of Proposition 1

This appendix provides the complete algebraic derivation underlying the three-step proof sketch given in Section 3.4.2. The proposition is restated for ease of reference, and each step is then proved in full.

Proposition A1 (Felt-Understanding Collapse, restated). *For any prior $p \in (0, 1)$ and cost structure satisfying the magnitude conditions of Section 3.2, there exists a threshold perturbation level $\pi^* = \max\{0, 1 - c_R/(u_R \cdot p)\}$ such that for all $\pi > \pi^*$, the unique Perfect Bayesian Equilibrium is the pooling equilibrium in which all agents of both groups play a_D . At $\pi = \pi^*$, type R is indifferent between a_R and a_D , and the pooling profile is selected under the standard tie-breaking convention, favoring the costless action. In this equilibrium, observed signals carry no information about types, and equilibrium felt understanding equals p while population behavior is uniformly dismissive regardless of the underlying recognition capacity distribution.*

Appendix A.1. Step 1: Type N Strictly Prefers a_D in Every Equilibrium

Consider an agent of type N. The agent's expected utility from playing a_R , conditional on any belief structure and channel perturbation level, has the following upper bound. The recognition payoff u_R is awarded only when the signal arrives undistorted at a type R receiver. The maximum probability of this joint event, taken over all possible posterior beliefs the agent might hold about the opposing group, is $1 \cdot p = p$, attained in the degenerate case where the agent believes the channel is undistorted and acts against an opposing group that is certainly playing the separating strategy. The identity-protection penalty u_I is incurred with certainty whenever type N plays a_R . The signaling cost c_N is also incurred with certainty whenever a_R is played by type N. Therefore:

$$EU_N(a_R) \leq u_R \cdot p - u_I - c_N. \quad (A1)$$

By the magnitude condition $u_I + c_N > u_R$ (the third condition stated in Section 3.2), and noting that $p \leq 1$, it follows that $u_R \cdot p \leq u_R < u_I + c_N$. Substituting into Equation (A1):

$$EU_N(a_R) \leq u_R \cdot p - u_I - c_N < 0. \quad (A2)$$

The expected utility from playing a_D for type N is $EU_N(a_D) = 0$, since the dismissive signal is costless and the identity-protection penalty is not incurred. Comparing:

$$EU_N(a_D) = 0 > EU_N(a_R). \quad (A3)$$

Therefore, type N strictly prefers a_D regardless of π , regardless of beliefs about the opposing group's type distribution, and regardless of the opposing group's actions. Type N plays a_D in every equilibrium. This completes Step 1.

Appendix A.2. Step 2: Type R's Best-Response Threshold

Given Step 1, the opposing population's mix of types and actions is fully determined by the type distribution: a fraction p of the opposing population are type R (who may play either action depending on the equilibrium under consideration), and a fraction $1 - p$ are type N (who play a_D by Step 1). Consider a type R agent contemplating whether to play

a_R or a_D , conditional on the opposing-population type R subgroup also playing a_R (i.e., conditional on a candidate separating equilibrium).

The expected utility of playing a_R for type R, under the candidate separating profile, decomposes by the joint event (channel transmits faithfully, receiver is type R):

$$EU_R(a_R) = P(\text{faithful transmission}) \cdot P(\text{receiver} = R) \cdot u_R - c_R. \quad (\text{A4})$$

The channel transmits faithfully with probability $1 - \pi$ by Equation (4) of the main text. The receiver is type R with probability p by the prior. The signaling cost c_R is incurred with certainty whenever type R plays a_R . The identity-protection penalty u_I does not apply because the agent is type R (by Equation (3) of the main text, u_I is conditioned on $\theta_i = N$). Therefore:

$$EU_R(a_R) = (1 - \pi) \cdot u_R \cdot p - c_R. \quad (\text{A5})$$

This recovers Equation (6) of the main text. The expected utility of playing a_D for type R is $EU_R(a_D) = 0$. Type R plays a_R as a best response to the separating profile if and only if $EU_R(a_R) > EU_R(a_D)$, which gives:

$$(1 - \pi) \cdot u_R \cdot p > c_R. \quad (\text{A6})$$

Rearranging Equation (A6) and solving for π :

$$\pi < 1 - c_R / (u_R \cdot p) \equiv \pi^*. \quad (\text{A7})$$

When the right-hand side of the threshold expression is negative (which occurs when $c_R / u_R \geq p$, equivalently when $u_R \cdot p \leq c_R$), the inequality has no solution in the admissible range $\pi \geq 0$, and the separating profile is unsustainable for type R regardless of perturbation. In this dead zone, $\pi^* = 0$ is set by convention, so that the threshold expression in the proposition becomes $\pi^* = \max\{0, 1 - c_R / (u_R \cdot p)\}$, matching the statement of Proposition 1. This completes Step 2.

Appendix A.3. Step 3: Uniqueness of the Pooling Equilibrium for $\pi \geq \pi^*$

By Step 1, type N plays a_D in every equilibrium. By Step 2, for $\pi > \pi^*$, the separating profile in which type R plays a_R is strictly not a best response (the type R best-response inequality (A6) holds strictly in the reverse direction). The only candidate equilibrium consistent with both facts is the pooling equilibrium in which both types of both groups play a_D . This profile is then verified to be an equilibrium, shown to be the unique Perfect Bayesian Equilibrium for $\pi > \pi^*$, and selected by tie-breaking at the boundary $\pi = \pi^*$.

Under the pooling profile, every agent sends s_D . Because the channel has full support for any $\pi \in (0, 1/2]$ by Section 3.3, the receiver observes both $\hat{s} = s_A$ and $\hat{s} = s_D$ with positive probability under any sender strategy profile. Therefore, Bayes' rule applies to every observed signal and pins down posterior beliefs uniquely. Under the pooling profile in which both types send s_D , the marginal probabilities of observing s_A and s_D are:

$$P(\hat{s} = s_A) = \pi, P(\hat{s} = s_D) = 1 - \pi. \quad (\text{A8})$$

Both observed signals are independent of the sender's type (since both types play the same action a_D in the pooling profile). By Bayes' rule, the posterior on $\theta = R$ conditional on either observation equals the prior:

$$\mu(\theta = R | \hat{s} = s_A) = \mu(\theta = R | \hat{s} = s_D) = p. \quad (\text{A9})$$

A unilateral deviation from a_D to a_R is then checked and shown to be unprofitable for a type R agent under these posteriors. The deviation costs c_R with certainty. The recognition payoff is earned with probability $(1 - \pi) \cdot p$ (channel transmits faithfully and receiver is type R) by the same calculation as in Step 2. Therefore:

$$EU_R(\text{deviate to } a_R) = (1 - \pi) \cdot u_R \cdot p - c_R. \quad (\text{A10})$$

For $\pi \geq \pi^*$, the weak threshold condition $(1 - \pi) \cdot u_R \cdot p \leq c_R$ holds; for $\pi > \pi^*$, the inequality is strict. The deviation payoff, therefore, satisfies:

$$EU_R(\text{deviate to } a_R) \leq 0 = EU_R(a_D). \quad (\text{A11})$$

Therefore, no type R agent has a strict incentive to deviate from a_D , and the pooling profile is a Perfect Bayesian Equilibrium for all $\pi \geq \pi^*$. For $\pi > \pi^*$, the threshold condition holds strictly, the deviation EU is strictly less than zero, and the pooling profile is the unique Perfect Bayesian Equilibrium. At the boundary $\pi = \pi^*$, Equation (A11) holds with equality: the type R agent is exactly indifferent between a_R and a_D , and both the separating profile and the pooling profile are equilibria. Under the standard tie-breaking convention, an indifferent agent plays the costless action; the pooling equilibrium is, therefore, selected at the boundary, and the collapse result extends to $\pi \geq \pi^*$. Note that no off-path beliefs need to be specified or refined because the full-support property of the channel ensures that every observable signal is on the equilibrium path with positive probability.

Equilibrium felt understanding in this pooling equilibrium, measured as the posterior $\mu(\theta = R | \hat{s})$ averaged over the equilibrium distribution of observed signals, equals $\pi \cdot p + (1 - \pi) \cdot p = p$, matching the statement of Proposition 1. Population behavior under the equilibrium is uniformly dismissive (every agent plays a_D) regardless of the underlying recognition-capacity distribution. This completes Step 3 and the proof of Proposition 1.

Three brief observations on the proof structure. First, the proof does not invoke any equilibrium refinement beyond Perfect Bayesian Equilibrium; the full-support property of the perturbed channel (Section 3.3) makes the Intuitive Criterion (Cho & Kreps, 1987) and related refinements unnecessary. Second, the threshold expression $\pi^* = 1 - c_R / (u_R \cdot p)$ is exact rather than asymptotic: it identifies the precise boundary at which the type R best response flips, not merely an approximation valid in some limit. Third, the dead-zone correction $\pi^* = \max\{0, \dots\}$ arises naturally from the requirement that $\pi \geq 0$; populations with sufficiently low prior recognition capacity (specifically, $p \leq c_R / u_R$) are in the pooling basin without any adversarial perturbation at all, and the threshold concept does not apply in the interesting sense in such populations.

Appendix B. The Adversary's Targeting Problem

The main text treats the adversary as a parameter of the environment (Section 3.1). This appendix makes explicit the constrained optimization that the targeting experiments of Section 5.5 numerically solve, so that the volume-versus-precision result can be read as a statement about a rational, budget-constrained attacker rather than only as a simulation outcome.

Fix the channel perturbation π and an injection budget $\gamma \in [0, 1]$, interpreted as the fraction of cross-group transmissions the adversary may corrupt at each step, and let M be the number of cross-group transmissions. A targeting policy τ assigns nonnegative weights to cross-group edges; the adversary injects exactly $\text{round}(\gamma \cdot M)$ fabricated dismissive signals per step, sampled without replacement with probability proportional to those weights. The three policies of Section 4.4 correspond to uniform weights, weights proportional to the receiving node's degree (central), and weights proportional to edge betweenness (bridge).

Holding γ fixed across policies isolates the allocation of a common budget from its size. The adversary's problem is then to choose $\tau \in \{\text{uniform, central, bridge}\}$ so as to maximize the induced long-run cross-group misrecognition $D(\pi, \gamma, \tau)$, defined operationally as one minus the long-run action rate averaged over the two groups, subject to that fixed budget.

Figure 5 evaluates $D(\pi, \gamma, \tau)$ on the stochastic block model and on a large scale-free network across the injection range, and so it numerically characterizes the solution of this problem. The maximand is approximately flat in τ : uniform targeting is weakly dominant at most injection rates, while central and bridge targeting are statistically indistinguishable from one another and never strictly dominate uniform. The argmax is, therefore, (weakly) uniform targeting on the topologies and budgets tested.

Two implications follow. First, a budget-constrained adversary that is approximately indifferent on the substantive margin will prefer the policy with the lowest intelligence and identification overhead—uniform volume—which rationalizes the empirically documented operational preference (Paul & Matthews, 2016; Fridman, 2023). Second, because the objective is approximately flat in τ , a fully strategic adversary endowed with its own utility function would reach the same qualitative targeting conclusion as the parametric adversary used in the main text, so the parametric reduction does not understate the threat. A complete three-player treatment, in which the adversary additionally chooses π and γ optimally against a defender, is left to future work (Section 6.4).

References

- Acemoglu, D., Ozdaglar, A., & Siderius, J. (2024). A model of online misinformation. *Review of Economic Studies*, 91(6), 3117–3150. [CrossRef]
- Balcaen, P., Du Bois, C., & Buts, C. (2021). The hybridisation of conflict: A prospect theoretic analysis. *Games*, 12(4), 81. [CrossRef]
- Baumann, F., Lorenz-Spreen, P., Sokolov, I. M., & Starnini, M. (2020). Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, 124(4), 048301. [CrossRef] [PubMed]
- Bennhold, K., & Solomon, E. (2022, December 2). How a German State helped Russia Complete Nord Stream 2. *The New York Times*. Available online: <https://www.nytimes.com/2022/12/02/world/europe/germany-russia-nord-stream-pipeline.html> (accessed on 21 May 2026).
- Boffey, D., & Oltermann, P. (2022, January 26). Germany's offer to Ukraine of 5000 helmets is 'joke', says Vitali Klitschko. *The Guardian*. Available online: <https://www.theguardian.com/world/2022/jan/26/russia-ukraine-germany-under-pressure-to-back-eu-military-training-mission-in-ukraine> (accessed on 21 May 2026).
- Bradshaw, S., & Howard, P. N. (2019). *The global disinformation order: 2019 global inventory of organised social media manipulation*. Working Paper No. 2019.3. Computational Propaganda Research Project, Oxford Internet Institute, University of Oxford.
- Campbell, T. H., & Kay, A. C. (2014). Solution aversion: On the relation between ideology and motivated disbelief. *Journal of Personality and Social Psychology*, 107(5), 809–824. [CrossRef] [PubMed]
- Chilton, S., Jones-Lee, M., Kiraly, F., Metcalf, H., & Pang, W. (2006). Dread risks. *Journal of Risk and Uncertainty*, 33(3), 165–182. [CrossRef]
- Cho, I.-K., & Kreps, D. M. (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2), 179–221. [CrossRef]
- Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6), 1431–1451. [CrossRef]
- Dennett, D. C. (1987). *The intentional stance*. MIT Press.
- Durrheim, K., & Schuld, M. (2025). Polarization on social media: Comparing the dynamics of interaction networks and language-based opinion distributions. *Political Psychology*, 46(6), 1601–1617. [CrossRef]
- Fairchild, R. (2026). Weaponized disinformation through dread: Russia, nuclear fear, and the geopolitical undermining of climate policy. *The Journal of Psychohistory*, 53(3), 187–209.
- Feng, Y., Ho, C.-J., & Tang, W. (2024). Rationality-robust information design: Bayesian persuasion under quantal response. In *Proceedings of the 2024 annual ACM-SIAM symposium on discrete algorithms (SODA)* (pp. 501–546). SIAM. [CrossRef]
- Frey, F. E., & Tropp, L. R. (2006). Being seen as individuals versus as group members: Extending research on metaperception to intergroup contexts. *Personality and Social Psychology Review*, 10(3), 265–280. [CrossRef] [PubMed]
- Fridman, O. (2023). Inherent strategic ambiguity between objectives and actions: Russia's 'information war'. *Defence Strategic Communications*, 12, 187–230. [CrossRef]

- Harvey, F. (2014, June 19). Russia ‘secretly working with environmentalists to oppose fracking’. *The Guardian*. Available online: <https://www.theguardian.com/environment/2014/jun/19/russia-secretly-working-with-environmentalists-to-oppose-fracking> (accessed on 21 May 2026).
- Hausken, K., Welburn, J. W., & Zhuang, J. (2024). A review of attacker–defender games and cyber security. *Games*, 15(4), 28. [CrossRef]
- Hennigs, R. (2021). Conflict prevention by Bayesian persuasion. *Journal of Public Economic Theory*, 23(4), 710–731. [CrossRef]
- Ioku, T., & Watamura, E. (2025). Further evidence for the role of felt understanding in intergroup relations: Japanese and Chinese relations in Japan. *Japanese Psychological Research*, 67(1), 87–97. [CrossRef]
- Jarvis, S., Deschenes, O., & Jha, A. (2022). The private and external costs of Germany’s nuclear phase-out. *Journal of the European Economic Association*, 20(3), 1311–1346. [CrossRef]
- Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4), 620–630. [CrossRef]
- Jobson, L., Wade, K. A., Rasor, S., Spearing, E., McEwen, C., & Fahmi, D. (2023). Associations between the misinformation effect, trauma exposure and symptoms of posttraumatic stress disorder and depression. *Memory*, 31(2), 179–191. [PubMed]
- Kamenica, E., & Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6), 2590–2615. [CrossRef]
- Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435), 1343–1370. [CrossRef]
- Kerman, T. T., Herings, P. J.-J., & Karos, D. (2024). Persuading sincere and strategic voters. *Journal of Public Economic Theory*, 26(1), e12671. [CrossRef]
- Kinderman, P., Dunbar, R., & Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, 89(2), 191–204. [CrossRef]
- Knight, B. (2011, March 15). *Off the grid*. DW. Available online: <https://www.dw.com/en/merkel-shuts-down-seven-nuclear-reactors/a-14912184> (accessed on 21 May 2026).
- Krishnan, A. (2024). *Fifth generation warfare: Dominating the human domain*. Routledge.
- Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32(2), 311–328. [CrossRef]
- Lei, L., Wu, Y., & Gao, S. (2026). Evolutionary game analysis of AI-generated disinformation governance on UGC platforms based on prospect theory. *Systems*, 14(4), 416. [CrossRef]
- Le Treust, M., & Tomala, T. (2016). Information design for strategic coordination of autonomous devices with non-aligned utilities. In *2016 54th annual Allerton conference on communication, control, and computing (Allerton)* (pp. 233–242). IEEE. [CrossRef]
- Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the “post-truth” era. *Journal of Applied Research in Memory and Cognition*, 6(4), 353–369. [CrossRef]
- Little, A. T. (2017). Propaganda and credulity. *Games and Economic Behavior*, 102, 224–232. [CrossRef]
- Livingstone, A. G., Fernández Rodríguez, L., & Rothers, A. (2020). “They just don’t understand us”: The role of felt understanding in intergroup relations. *Journal of Personality and Social Psychology*, 119(3), 633–656. [CrossRef] [PubMed]
- Livingstone, A. G., Spears, R., Manstead, A. S. R., & Bruder, M. (2009). Illegitimacy and identity threat in (inter)action: Predicting intergroup orientations among minority group members. *British Journal of Social Psychology*, 48(4), 755–775. [CrossRef] [PubMed]
- Lu, H.-C., & Lee, H.-W. (2025). Agents of discord: Modeling the impact of political bots on opinion polarization in social networks. *Social Science Computer Review*, 43(4), 750–772.
- Paul, C., & Matthews, M. (2016). *The Russian “firehose of falsehood” propaganda model: Why it might work and options to counter it*. Perspective No. PE-198-OSD. RAND Corporation. [CrossRef]
- Pomerantsev, P., & Weiss, M. (2014). *The menace of unreality: How the Kremlin weaponizes information, culture and money*. Institute of Modern Russia.
- Reis, H. T., Lemay, E. P., Jr., & Finkenauer, C. (2017). Toward understanding understanding: The importance of feeling understood in relationships. *Social and Personality Psychology Compass*, 11(3), e12308. [CrossRef]
- Rogers, C. R. (1995). *A way of being*. Houghton Mifflin.
- Seligman, M. E. P. (1975). *Helplessness: On depression, development, and death*. W. H. Freeman.
- Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research*, 177(3), 1333–1352. [CrossRef]
- Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3), 355–374. [CrossRef]
- Stamp, D. (2011, March 15). Germany suspends deal to extend nuclear plants’ life. *Reuters*.
- Tsintsaris, D., Tsompanoglou, M., & Ioannidis, E. (2024). Dynamics of social influence and knowledge in networks: Sociophysics models and applications in social trading, behavioral finance and business. *Mathematics*, 12(8), 1141. [CrossRef]
- U.S. Department of State, Global Engagement Center. (2020, August). *Pillars of Russia’s disinformation and propaganda ecosystem*. GEC Special Report. U.S. Department of State, Global Engagement Center.
- U.S. House of Representatives, Committee on Science, Space, and Technology. (2018). *Russian attempts to influence U.S. domestic energy markets by exploiting social media*. Majority Staff Report. U.S. House of Representatives, Committee on Science, Space, and Technology.

- Vendeville, A., & Diaz-Diaz, F. (2025). Modeling echo chamber effects in signed networks. *Physical Review E*, 111(2), 024302. [[CrossRef](#)] [[PubMed](#)]
- Vorauer, J. D., Main, K. J., & O'Connell, G. B. (1998). How do individuals expect to be viewed by members of lower status groups? Content and implications of meta-stereotypes. *Journal of Personality and Social Psychology*, 75(4), 917–937. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.