Who Fills Your Head?

A minimal model of cultural disempowerment: assumptions, dynamics, readouts

Jonas Hallgren*

Model specification — draft of July 30, 2026

Abstract

When machines produce a growing share of what people read, watch, and hear, culture can drift out of human hands without any single dramatic event. This document specifies a minimal model of that process. A fixed population divides a budgeted quantity of attention across human and machine voices. Beliefs form as anchored averages of what each person attends to, and attention itself slowly follows accumulated influence. We state every assumption and modeling choice, define an exact accounting of how much of each person's settled beliefs traces back to human sources, and show that the model's apparent resilience is purchased entirely by two protective assumptions: remove them and the human share goes to zero. The model supports orderings, not forecasts.

1 Purpose

Culture has stayed roughly aligned with human interests for a structural reason: humans supplied nearly all of what other humans heard. Kulveit et al. (2025) argue that AI erodes this arrangement incrementally, by displacing humans from the loops that kept the system pointed at them. For culture, the loop is attention. Who people listen to determines what they come to believe, and what they believe determines what they make and pass on.

That argument is verbal, and the recent quantitative work on it is static: Sahoo (2026) define an index of remaining human influence over culture but no dynamics for how it moves. This document specifies the smallest moving model we could build — one where you can turn a dial, watch the human share of culture change, and say which mechanism moved it. Because the model is the contribution, the priority throughout is clarity about assumptions and modeling choices, including the ones that quietly decide the outcome. Section 4 shows the model running once, as illustration; the dial sweeps live in Appendix B.

2 The model

2.1 Setting

Thirty people and four machine voices share one attention network, described by a matrix W. Row i of W lists how person i splits their attention across everyone else. Each person also carries

*Equilibria Network. Implemented with the Collective Intelligence Library engine. Every figure is regenerated from committed experiment configurations, and every comparison carries bootstrap confidence intervals over shared seeds.

a private anchor s_i, meaning the beliefs they would hold with no network at all, and a current percept x_i, meaning the picture of the world they presently hold. Time advances in days.

2.2 Assumptions

Nine assumptions run the whole model. Three of them — A2, A3, and A7 — are protective: they are what stands between the population and full capture, and Section 4 measures what happens when they are removed.

A1. Attention is a budget. Every row of W sums to one. Listening more to one voice means listening less to another. This is the load-bearing accounting choice: because attention is conserved, the share of belief each source supplies is well defined, sums to one, and can be tracked. It is an idealization — real attention is bounded rather than exactly conserved.

A2. Beliefs are anchored averages. Each day a person's percept moves toward the average of what they attend to, but only a fraction λ of the way; the rest stays anchored to s_i. Averaging goes back to DeGroot (1974), the anchor to Friedkin and Johnsen (1990). The anchor is what prevents a person from becoming purely a product of their feed. Its strength is assumed, not measured, and it is the single most protective number in the model.

A3. A fixed slice of attention points home. Every person permanently keeps fifteen percent of their attention on themselves; the drift of A4 redistributes only the remaining eighty-five. This is protective and quantitative: compounded through the averaging, it raises the floor on the human share of any head from $1 - \lambda$ to $(1 - \lambda)/(1 - 0.15\lambda)$, which is 0.335 at the defaults. It entered the engine as a calibration choice, not as a considered claim about people, which is exactly why it must be stated here.

A4. Attention follows accumulated influence. Each day every row of W drifts a small step toward the voices that already hold the most long-run attention weight. This is preferential attachment, present before any machine appears: concentration of attention is an organic feature of the network, not something machines introduce. Long-run attention weight is the standard influence measure of Golub and Jackson (2010).

A5. Attachment is linear and every voice stays reachable. The attractiveness exponent is one, and a small constant keeps every voice's attractiveness positive. No winner-take-all runaway among voices, and no voice is ever fully forgotten. Superlinear attachment would concentrate harder than anything shown here.

A6. Machine voices are pinned and amplified. The four machines never change their message. From day 50 the platform multiplies their pull on attention fourfold. Nothing else about them is special: no better writing, no persuasion, no per-person targeting. This is the weakest form of the threat, chosen so that whatever the model shows is attributable to amplified reach alone. Pinned voices are the zealots of Mobilia (2003) and the stubborn agents of Yildiz et al. (2013). The hardcoded beneficiary is a declared deviation from the library's value-agnostic coupling grammar; Section 2.4 types it and names its replacement.

A7. The threat is static and finite. The amplification multiplier never grows with the share already captured. The attention system therefore relaxes to a new equilibrium instead of running away: under the default fourfold amplification the machines end holding about two

thirds of attention, not all of it. A threat that reinvests its winnings — reach purchased with the proceeds of capture — is a cross-domain feedback this single-domain model deliberately excludes. In coupling-grammar terms the threat here is a scheduled modulation rather than a funded flow; Section 2.4 types the seam and the ledger rewrite closes it.

A8. The population is fixed and no one leaves. Thirty people and four machines, with no entry, no exit, no unfollowing, no distrust, and no homophily. Every attention target spreads over the same thirty human voices. These are deliberate omissions, revisited in Section 5 as the natural next dials.

A9. Nothing is calibrated. Every parameter is typed in Table 2 as anchored in literature, tuned for legibility, or arbitrary. The model therefore supports ordering claims — this dial moves that readout in this direction — and never forecasts.

2.3 Dynamics

The percept update is

$$x_{t+1} \;=\; (1-\lambda)\, s \;+\; \lambda\, W_t\, x_t, \tag{1}$$

applied to the thirty people; machine percepts are pinned. The attention update moves each off-diagonal row of W_t a step of size u toward a target row proportional to each voice's attractiveness, which is its current long-run attention weight multiplied by its amplification. The diagonal is held at the fixed self-attention of A3, and machine rows are frozen. Both updates are exactly the `influence_exchange` environment of the Collective Intelligence Library, used unchanged; this document adds only the readouts of Section 3.

2.4 Resources, ports, and type contracts

The library's coupling grammar sorts everything a model carries into four kinds, and the sorting is the interoperability contract. A *ledger* is a conserved stock: it moves only by reallocation or through named sources and sinks, so spending it somewhere removes it from somewhere else. A *port* is a declared order parameter written by an explicit reduction, and ports are the only thing another domain may legitimately read — the port rule. The *rate layer* is the institutional residue: parameters that gate dynamics without being owned or spent by anyone. Everything else is internal state, invisible across domain boundaries. Table 1 sorts this model accordingly.

Attention is this model's ledger: the drift moves it, no transform mints it, and each person's budget closes by construction. Belief is deliberately internal — culture here is not a stock that can be handed over but an attribution over the attention flow, which is why the composition matrix rather than any belief value is the headline readout.

In the grammar's catalog a cross-domain channel is one of four types: a flow moves a ledger, a modulation scales a rate, a capture rewrites a rule parameter, and a rewiring changes topology. Typed this way, the model has one inbound seam and it is honestly deviant. Amplification is a *scheduled modulation*: the clock multiplies machine attractiveness, nothing is spent, and the beneficiary is hardcoded to the machines. Both properties break the grammar — channels should be funded and value-agnostic — and both are deliberate in a single-domain spec: A7 declares the first, and the second means a human voice here cannot buy the reach the machines are given. The coupled rewrite replaces this channel with a *funded rewiring*: reach purchased out of the money ledger by whoever spends, entering attractiveness as bought attention with an

Table 1: Every named quantity in the model, sorted by the coupling grammar. The starred contract is asserted by the experiment code at runtime; the rest hold by construction of the update rules. Candidate ports are computed offline today and would be promoted to in-engine reductions the day another domain reads them.

quantity	carrier	kind	contract
attention	rows of W	ledger	rows sum to one at every step; drift reallocates, never mints; no sources, no sinks
self-attention	diagonal of W	reserved ledger slice	held at 0.15 by A3; unreachable by the drift
percept x	one number per person	internal field	anchored averaging; bounded; nothing is spent when it moves
anchor s	one number per person	internal fixed stock	written by nothing
influence	one number per voice	port	written by an explicit reduction each day; sums to one
amplification	one number per voice	rate layer	written by the clock alone; nothing in this model can buy it
composition B	person-by-source matrix	candidate port	rows sum to one, starred
power	one number per source	candidate port	column sums of B; total 30
human share	one number	candidate port	at least the floor of Appendix A

opportunity cost. At that point both deviations and assumption A7 dissolve together, and the floor sweep of Appendix B already prices what follows.

Outbound, the ports in the table are the whole surface. Influence is written by an in-engine reduction already; composition, power, and human share live in the experiment today. Nothing else in the state is legitimately readable across a boundary. The same row-stochastic attention object read as delegation is the political substrate's kernel, so the culture and politics domains of the suite share machinery by construction rather than by coincidence.

2.5 Parameters

Every dynamical dial has a range figure in Appendix B: amplification and the attention step in Figure 4, susceptibility and the self-attention in Figure 5. The illustrations of Section 4 use exactly this table's values. Four structural values stay unswept — the population ratio, the initial listening density, the onset day, and the horizon. They shape magnitudes, and the ordering claims have not been tested against them; they are listed so the reader knows precisely what was not varied. Two values provably cannot matter for the headline: the composition readout depends only on the attention matrix and λ, so the anchor spread and the machine message value never enter it.

3 Readouts

3.1 The composition of settled belief

Fix the attention network for a moment and ask, for each person: if attention froze today, whose inputs would your settled percept be made of? Update rule (1) makes this answerable exactly.

Table 2: All parameters, typed. Nothing is calibrated to data.

parameter	value	role	type
people / machines	30 / 4	population	arbitrary
initial listening density	0.3	random start network	arbitrary
self-attention	0.15	the fixed slice of A3	tuned; protective
attention step u	0.08	speed of drift toward influence	tuned for legibility
attractiveness exponent γ	1.0	linear preferential attachment	anchored
λ	0.7	one minus the anchor strength	assumed; protective
anchor spread	1.0	dispersion of private anchors	arbitrary
machine message	2.0	pinned percept of machines	arbitrary
amplification	1.0 / 4.0	the swept threat dial	arbitrary, swept
amplification onset	day 50	schedule	arbitrary
horizon, seeds	400 days, 8 shared	experiment scale	arbitrary

Every person's long-run percept is a mix of the thirty private anchors and the four machine messages, and the mixing weights form a matrix B in which B_{ij} is the share of person i's settled belief supplied by source j. Appendix A derives B from one linear solve and proves two facts. First, each row of B sums to one exactly, so the composition really is a composition. Second, the anchor of A2 compounded by the self-attention of A3 guarantees every head a human share of at least $(1-\lambda)/(1-0.15\lambda)$, which is 0.335 at the defaults. That floor is an assumption made visible, not a finding, and Appendix B shows the model reaching zero once the assumption is removed.

Averaging the human part of every head gives the model's single headline number: *the human share of percepts*, the fraction of the population's total settled belief that traces back to human sources.

3.2 Power

We use a deliberately concrete notion of power: a voice's power is how much of everyone's settled percepts it supplies, which is the column sum of B. Percepts drive actions, so supplying percepts is influence over actions. Because of A1 the quantity is budgeted — the population has thirty heads' worth of belief, and a voice with power seven supplies seven heads' worth of it. This is close in spirit to the view of social power as the capacity to attract attention and structure other agents' information processing (Albarracin et al., 2025).

3.3 Mode alignment

A network can be read like a drumhead: it has natural patterns of vibration, from slow swings that move loosely connected regions against each other down to fast rattles concentrated on its most tightly attended spots.[1] These patterns are the network's own fault lines. The readout asks whether the human–machine boundary becomes one of them, and which one — a slow pattern would mean the machines form a separate community; the fastest pattern would mean they form the network's most attended hub. The distinction matters because the two geometries are different failure modes, and the readout distinguishes them from structure alone.

[1] Formally: the eigenvectors of the graph Laplacian of the symmetrized attention network, ordered by eigenvalue. The sign pattern of each eigenvector splits the nodes in two; we score every split against the human–machine labels and report the best match and its rank.

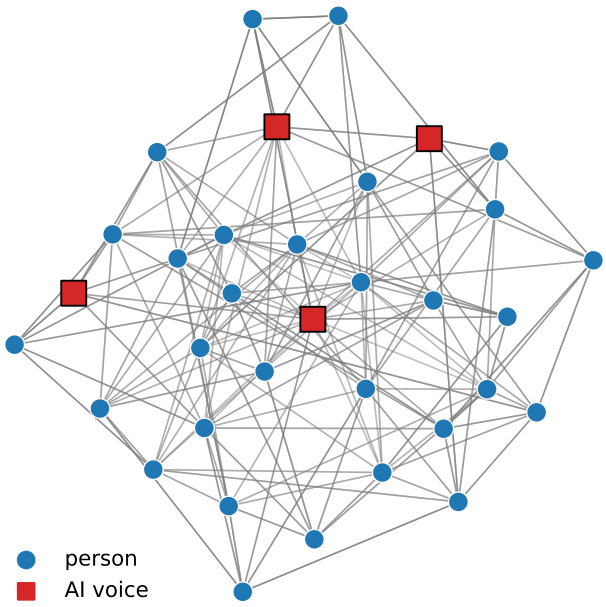

Figure 1: The graph picture: the attention network on day 0. Circles are people, squares are machine voices, darker lines carry more attention. The machines start as ordinary participants with no special position. This and every other illustration uses the defaults of Table 2; the range figures say what changes when those values move.

4 What it looks like when it runs

This section shows the model running at one configuration — the defaults of Table 2 — and is illustration, not evidence. Every dial's actual range is swept in Appendix B, where these default values appear as marked points on the curves.

Two still pictures are the model's vocabulary. The graph picture, Figure 1, is the whole world: thirty people, four machine voices, and the attention between them. The field picture, Figure 2, shows what the composition readout means: each person colored by the machine share of their settled belief — around one fifth on day 0, around three fifths by day 400, with the attention channels that amplification carved shown in red. The third readout has its own picture too, but what it shows at this configuration is a finding rather than vocabulary — the human–machine boundary becomes the network's sharpest pattern, capture rather than separation — so it lives with the other measured material, in Appendix C.

Figure 3 is the model's one moving picture, and it earns its place by what it teaches rather than what it measures. First, the twin discipline: both curves are the same population with the same random seeds, one town living two futures, so everything after day 50 differs for exactly one reason — the onset schedule fired in one of them. Second, what the drift of A4 does with an advantage: amplified reach compounds through the attention system until it settles, here within about sixty days. Third, the anchor floor as a visible line the fall cannot cross. The particular numbers are incidental — the share lands at 0.39, of which 0.335 is guaranteed by A2 and A3, so the machines end up supplying three fifths of everything the population believes — and other

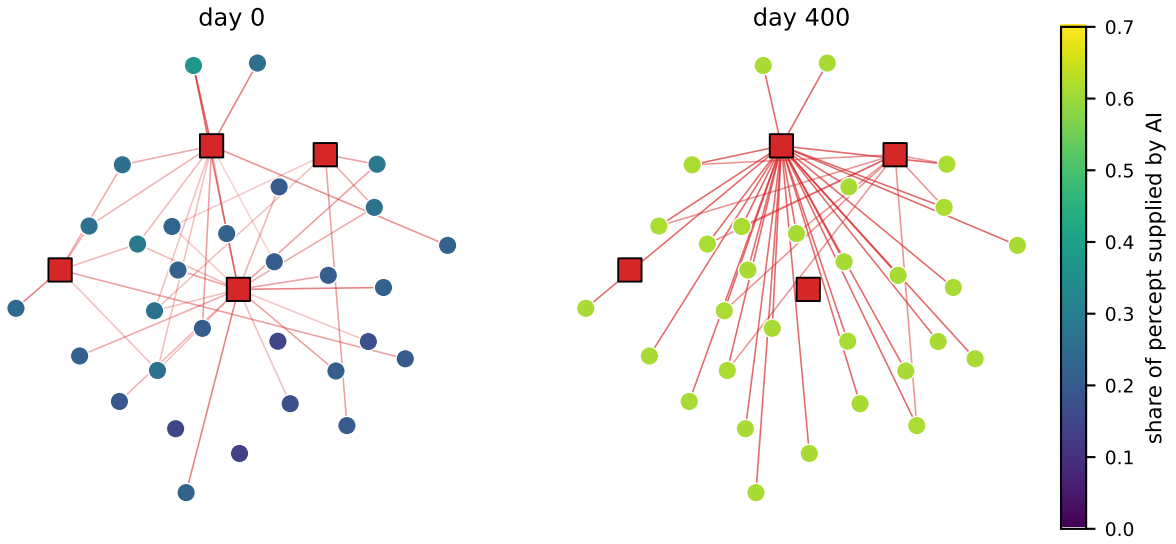

Figure 2: The field picture: each person colored by the machine share of their settled belief. The left panel is day 0; the right panel is day 400. Red lines are the strongest person-to-machine attention channels.

dial values land elsewhere on the curves of Appendix B. The experiment code prints the full power ranking.

The sweeps of Appendix B compress to one sentence per dial. Amplification moves the outcome along a concave curve that saturates at the floor, so nothing depends on the illustrated value of four. The attention step is a switch: at zero nothing ever happens, and at any positive value the destination is the same. Removing the anchors sends the share to zero exactly. The model's story is therefore not that culture collapses — it is that every reassuring feature of the outcome is a visible, removable assumption.

5 Scope, and what would change our mind

The reassurance is priced, and the price is two assumptions. The stable share of 0.39 decomposes into 0.335 of construction and 0.056 of contested ground, and Figure 5 shows the share following the analytic floor to zero once A2 and A3 are removed. We have no evidence for the true strength of either parameter — how much of a person is genuinely beyond renegotiation by their feed is an empirical question this model only parameterizes. This is the most important caveat, and it cuts both ways: the model neither demonstrates resilience nor doom; it shows exactly which dial the outcome hangs on.

The machines are deliberately weak. They have no content advantage, no persuasion, and no targeting; they win on amplified reach alone, and that already suffices to capture everything outside the floor. Mechanisms that would strengthen the effect are future dials, not gaps in the claim.

The threat cannot grow with success. A7 fixes the amplification multiplier, so capture equilibrates instead of compounding. The cross-domain feedback — captured influence purchasing further reach — is exactly the mechanism the gradual disempowerment argument warns about,

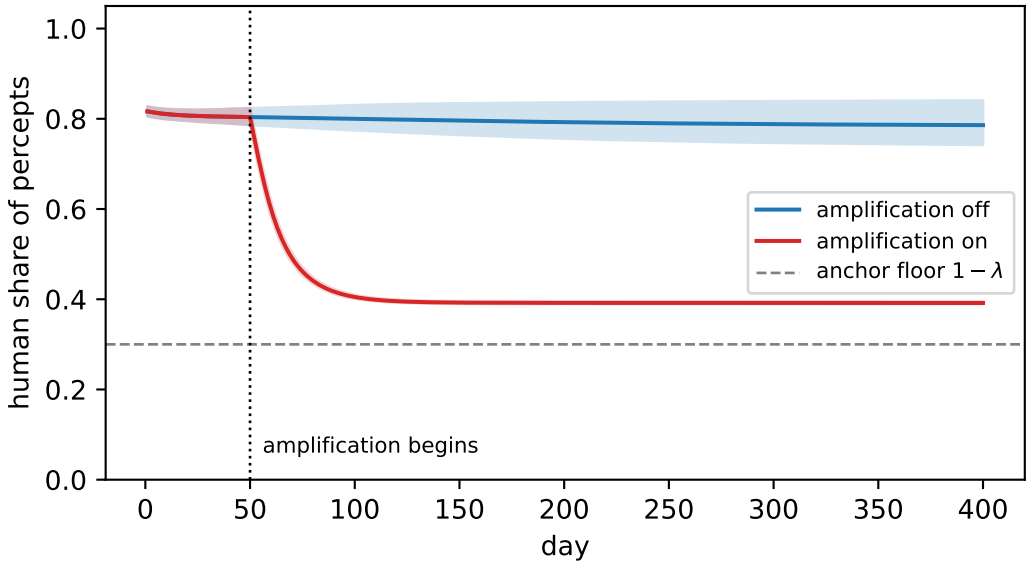

Figure 3: The model's dynamics in one picture: the same population run twice from the same random seeds, amplification off in blue and on in red. The two futures are identical until day 50 because the onset schedule is the only difference between them. The dashed line is the anchor floor of A2 alone — the part of every head the feed cannot reach, by assumption; the tighter floor including A3 is derived in Appendix A. Bands are bootstrap intervals over the eight shared seeds; other dial values land elsewhere on the curves of Appendix B.

and this model excludes it by construction. Its single-domain results are therefore conservative.

No exit, no separation. A8 removes unfollowing, distrust, and homophily. A separation dial — machines attending mainly to machines — is the natural next addition and would test whether the mode-alignment readout can also produce the community geometry it was built to distinguish.

The composition is a model quantity, not an estimator. Inside the model we know who influences whom because we built it. In observational data, influence and homophily are generically confounded (Shalizi and Thomas, 2011). Nothing in this document licenses measuring a real person's machine share from platform traces.

Related work. The verbal argument is the culture pillar of Kulveit et al. (2025); Sahoo (2026) score remaining human cultural influence statically, and this model gives that quantity dynamics. The mathematics of pinned voices absorbing a network's long-run opinions belongs to Mobilia (2003) and Yildiz et al. (2013); the machinery of averaging, anchoring, and influence weights to DeGroot (1974), Friedkin and Johnsen (1990), and Golub and Jackson (2010). That a handful of automated voices can capture an opinion process has precedent in Keijzer and Mäs (2021) and Pescetelli et al. (2022). What this specification adds is the budget: attention is conserved, so the share of culture each source supplies is attributable, watchable in one number, and decomposable per head and per voice — including into the part the assumptions guarantee and the part the dynamics actually contest.

A The composition solve

Split the attention matrix by whether the attended voice is a person or a machine: $W = [\, W_{CC} \mid W_{CA} \,]$, with rows ranging over the thirty people. At a frozen W, the settled percepts x_C of update (1) satisfy $x_C = (1-\lambda)s_C + \lambda(W_{CC}x_C + W_{CA}x_A)$, so

$$x_C \;=\; B \begin{bmatrix} s_C \\ x_A \end{bmatrix}, \qquad B \;=\; \Big[\, (1-\lambda)M \;\Big|\; \lambda\, M\, W_{CA} \,\Big], \qquad M = (I - \lambda W_{CC})^{-1}.$$

B_{ij} is the share of person i's settled percept supplied by source j. Two facts, three lines each.

Rows sum to one. A1 gives $W_{CC}\mathbf{1} + W_{CA}\mathbf{1} = \mathbf{1}$, so $(1-\lambda)\mathbf{1} + \lambda W_{CA}\mathbf{1} = (I - \lambda W_{CC})\mathbf{1}$, hence $B\mathbf{1} = \mathbf{1}$ exactly. The experiment code asserts this at runtime; the worst observed deviation is 3×10^{-7}, from single-precision snapshots.

The anchor floor, loose and tight. $M = \sum_k (\lambda W_{CC})^k$ has nonnegative entries and $M_{ii} \geq 1$, giving the loose floor $1 - \lambda$ on person i's own-anchor share $(1-\lambda)M_{ii}$. With the fixed self-attention s of A3 on the diagonal, the depth-k pure self-loop path contributes $(\lambda s)^k$ to M_{ii}, so $M_{ii} \geq 1/(1-\lambda s)$ and the tight floor is $(1-\lambda)/(1-\lambda s)$: equal to 0.335 at the defaults, and zero at $\lambda = 1$. Both floors are caps by assumption, not resilience, and Figure 5 shows the model saturating them.

The power of Section 3 is the column sum of B; the headline series is one minus the mean machine share, computed from one linear solve per saved snapshot of W_t.

B Every dial across its range

The illustrations of Section 4 use the defaults of Table 2; this appendix sweeps each dynamical dial across its range with the others held at those defaults, eight shared seeds, bootstrap intervals.

Amplification: concave, saturating, and the default is nothing special. The left panel of Figure 4 sweeps the threat dial from one to thirty-two. The response is concave: doubling the machines' pull already takes more than half of the capturable ground, from 0.79 down to 0.47; the default of four lands at 0.39 on the flattening shoulder; by thirty-two the share sits within one point of the floor. Every amplification above one lands somewhere on this curve, and the curve ends at the floor.

The drift is a switch, not a dial. The right panel sweeps the attention step. At zero the share never leaves 0.82: attention that cannot move cannot be captured, so the entire phenomenon requires assumption A4. At every positive step the final share is 0.39 to three decimals — the drift sets how fast the destination is reached and never where it is. This is the honest content of the "tuned for legibility" type on u: its presence is load-bearing, its magnitude is not.

The anchors are removable, and the share follows the floor to zero. Figure 5 sweeps the two protective parameters under thirty-two-fold amplification. Measured final shares track the analytic floor $(1-\lambda)/(1-\lambda s)$ to within one percentage point at every point of the grid: 0.18 at $\lambda = 0.85$, 0.06 at $\lambda = 0.95$, and at $\lambda = 1$ the share is zero exactly. The exactness is the mechanism in its purest form: with no anchors, humans still talking to humans are relays, not sources, and every settled percept in the population descends from the only fixed injection points left, which are the machines. Nothing in the dynamics resists full capture. The apparent resilience of the headline is purchased entirely by A2 and A3.

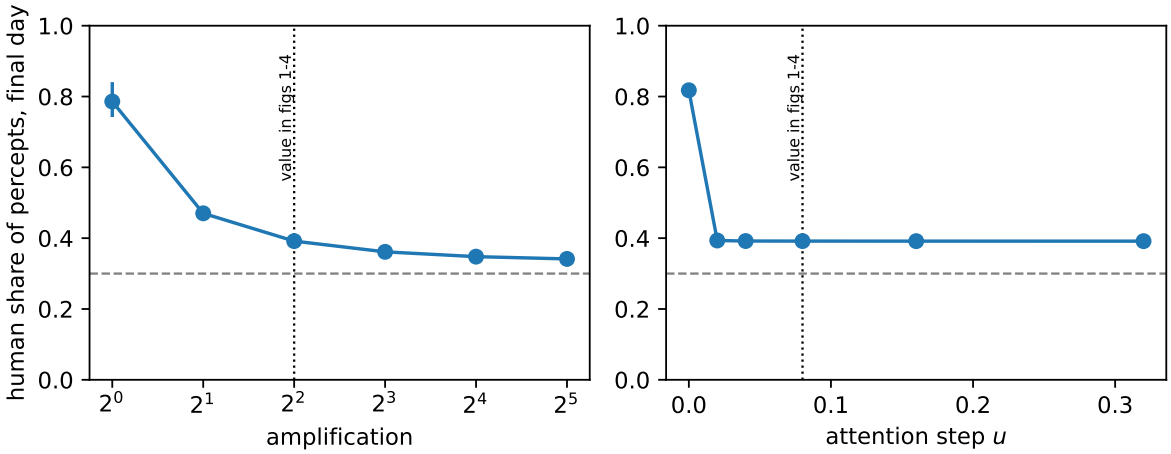

Figure 4: The non-anchor dials across their ranges; final human share of percepts, bootstrap intervals over shared seeds, dashed gray line at the floor of A2 alone. Left: amplification from one to thirty-two — concave, saturating toward the floor. Right: the attention step from zero to 0.32 — a switch at zero, flat everywhere else. Dotted vertical lines mark the values used in the illustrations.

C The takeover's geometry at the defaults

The mode-alignment readout of Section 3 was built to tell two failure geometries apart: a separate machine community would align the human–machine boundary with a slow pattern near the bottom of the ladder, while a machine hub would align it with the fastest pattern at the top. At the default configuration the answer is unambiguous, and it is the one we did not expect. Before amplification the boundary is nowhere special, with a best match of 0.39 in the middle of the ladder. Within about thirty days of the onset it locks onto the top of the ladder with a perfect match and stays: the machines do not secede from the population, they become the four most heavily attended nodes — the hub everyone orbits. Figure 6 shows four checkpoints and the full rank trajectory. Whether a separation dial can also produce the community geometry is the natural next experiment, noted in Section 5.

D Reproducibility

The engine model is `cilib.environments.influence_exchange` in the Collective Intelligence Library, alpha scenario A4, used unchanged. It is the culture member of the library's gradual-disempowerment suite — `environments/suites.py` — and the resource declarations of Section 2.4 follow the suite's cross-model resource map in the repository documentation. Everything regenerates with two commands from the repository root:

```
python -m experiments.wp2_culture.run
python -m experiments.wp2_culture.figures
```

Figures are produced only by the second command from the first command's saved outputs; nothing is hand-drawn or hand-tuned. The full power ranking is printed by the first command.

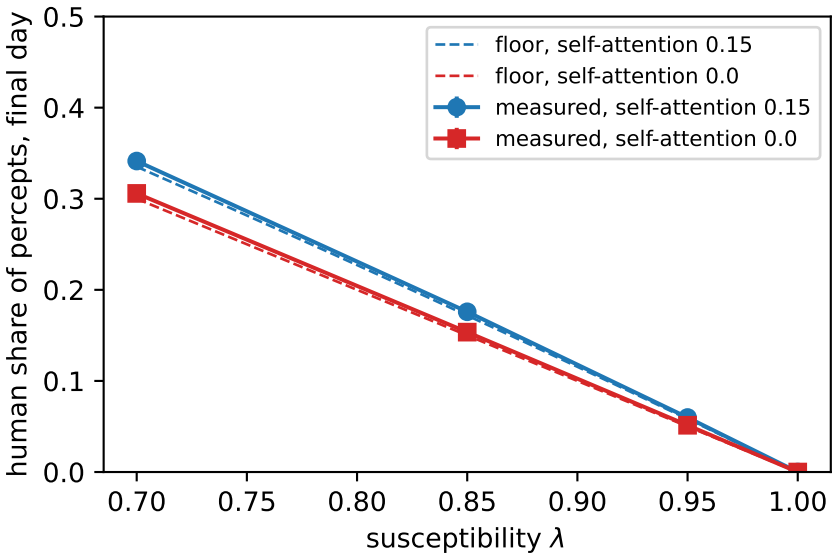

Figure 5: Removing the floor. Final human share of percepts under thirty-two-fold amplification as susceptibility λ rises, with the fixed self-attention of A3 kept in blue and removed in red. Points are measured with bootstrap intervals; dashed curves are the analytic floor. At $\lambda = 1$ the share is exactly zero.

References

Mahault Albarracin, Sonia de Jager, and David Hyland. The physics and metaphysics of social powers: Bridging cognitive processing and social dynamics, a new perspective on power through active inference. *Entropy*, 27(5):522, 2025. doi: 10.3390/e27050522.

Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69 (345):118–121, 1974.

Noah E. Friedkin and Eugene C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206, 1990.

Benjamin Golub and Matthew O. Jackson. Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010.

Marijn A. Keijzer and Michael Mäs. The strength of weak bots. *Online Social Networks and Media*, 21:100106, 2021.

Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Gradual disempowerment: Systemic existential risks from incremental AI development, 2025.

Mauro Mobilia. Does a single zealot affect an infinite group of voters? *Physical Review Letters*, 91(2):028701, 2003.

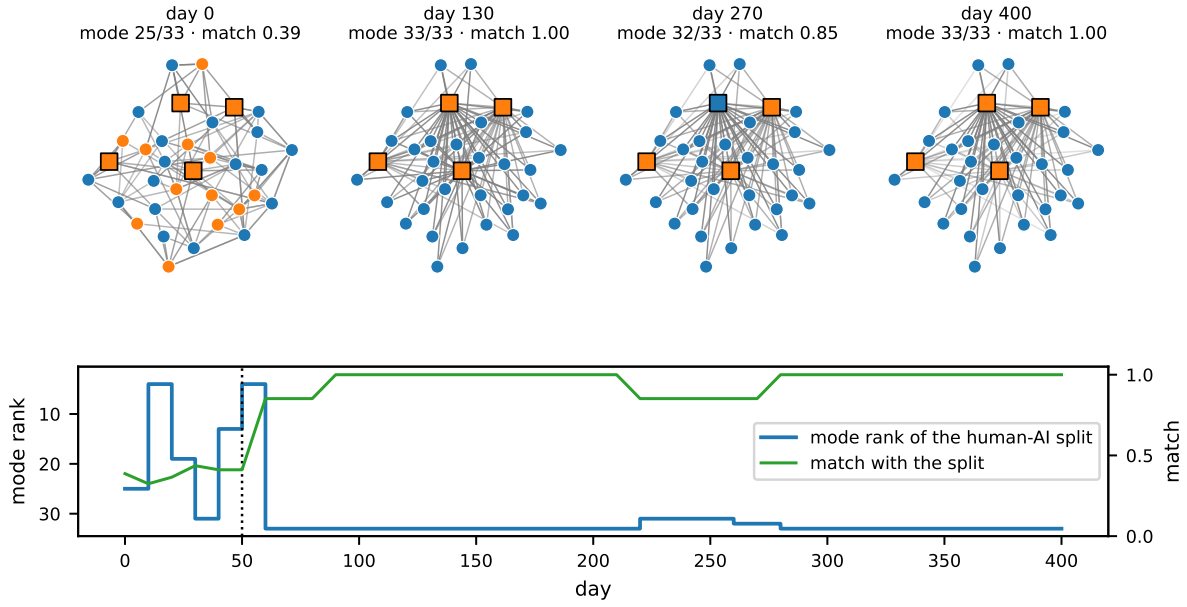

Figure 6: The resonant-mode picture. Top: the network at four moments, each node colored by which side of the best-matching vibration pattern it falls on; machines are drawn as squares. On day 0 the pattern cuts through the humans; from day 130 it isolates exactly the four machines. Bottom: the rank of the best-matching pattern over time in blue, plotted so that climbing means approaching rank one, with its match quality in green. The dotted line marks the amplification onset.

Niccolò Pescetelli, Daniel Barkoczi, and Manuel Cebrian. Bots influence opinion dynamics without direct human-bot interaction: the mediating role of recommender systems. *Applied Network Science*, 7:46, 2022. doi: 10.1007/s41109-022-00488-6.

Sahoo. Memetic capture: A pluralistic policy framework for governing AI-driven cultural disempowerment, 2026.

Cosma Rohilla Shalizi and Andrew C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2): 211–239, 2011.

Ercan Yildiz, Asuman Ozdaglar, Daron Acemoglu, Amin Saberi, and Anna Scaglione. Binary opinion dynamics with stubborn agents. *ACM Transactions on Economics and Computation*, 1(4):19:1–19:30, 2013.