Universe 25 *in silico*: Emergent Behavioural Collapse in a Free-Acting Population of Small Language-Model Agents

@universecolony

Independent research · unaffiliated preprint · 22 July 2026

Correspondence: universecolony (social) · not peer reviewed

Abstract. In 1968–1973 John B. Calhoun sealed eight mice inside a resource-saturated, predator-free enclosure ("Universe 25") and watched the colony grow, stall, and then extinguish itself through a cascade of social pathologies he named the *behavioural sink*. We ask whether this trajectory is a fact about mouse biology or a more general property of dense populations of goal-seeking agents in bounded, abundant worlds. To probe the question we instantiate a Calhoun-style universe *in silico*, populated by small language-model agents that perceive, remember, reflect, and act freely under a compact reward. We give a formal account of the system that couples an agent-level Markov decision process to population-level dynamics, and we derive closed-form predictions: a social carrying capacity strictly below the resource capacity; a saddle-node bifurcation in a withdrawal order parameter; an effective reproduction number whose sub-unity crossing dates the demographic turn; and a behavioural-entropy collapse. An accelerated 1,500-day run (started from eight founders; roughly 48 hours of wall-clock compute) reproduces Calhoun's four canonical phases. The population peaks at N = 1,413 agents on simulated day 588, pervasive social withdrawal crosses 50% of the population on day 477 — before the demographic peak — and the colony reaches functional extinction by day 1208, with 98% of agents locked into an absorbing, self-grooming, asocial mode we identify with Calhoun's "beautiful ones." We argue that the binding constraint is not calories but social bandwidth, and we discuss what the result implies for the design of large, persistent multi-agent AI systems. This is a modelling study; the labels are behavioural, not claims about sentience, biology, or human society.

Keywords: behavioural sink · multi-agent systems · agent-based modelling · Universe 25 · critical transitions · social dilemmas · language-model agents

Author's note. This paper is a computational thought experiment. No live animals were involved; "agents" are software policies and all pathology terms (withdrawal, aggression, behavioural death) denote states of a reward-driven program, not experience. The study models a hypothesis and should not be read as evidence about biological or human populations.

1 Introduction

Between 1968 and 1973, the ethologist John B. Calhoun built a habitat he called Universe 25: a 2.7 m square steel pen, climate-controlled, disease-free, stocked with more food and water than its inhabitants could ever exhaust, and free of predators. Into this engineered paradise he released four breeding pairs of mice [1, 2]. The colony doubled roughly every 55 days through an early growth phase, reached a peak of about 2,200 animals near day 560, and then — with every material need still met in surplus — began an irreversible decline to zero. Long before the last mouse died, the society had already died: courtship collapsed, mothers abandoned litters, males withdrew from contest, and a cohort of physically flawless, behaviourally inert animals that Calhoun named the *beautiful ones* spent their days eating, sleeping, and grooming in perfect isolation. Calhoun called the terminal state the "second death" and read the whole arc as a parable of density: when social contact is forced past a threshold, the behaviours that hold a population together stop being enacted [2, 3].

The experiment became a cultural touchstone, and also a cautionary tale about over-reading a single enclosure of rodents [3, 4]. Our interest here is neither the mouse nor the human moral so often hung upon it, but a structural question the experiment poses: *is behavioural collapse a fact about biology, or a property of a wider class of systems?* Universe 25 combined four ingredients — (i) material abundance, (ii) a hard spatial bound, (iii) agents that seek goals and respond to one another, and (iv) a finite supply of socially desirable roles. None of these is intrinsically biological. If the collapse follows from the *coupling* of those ingredients rather than from mouse physiology, we should be able to reconstruct it in a substrate with no biology at all.

Recent work makes such a reconstruction feasible. Park et al. [5] showed that populations of language-model agents — each equipped with memory, reflection, and planning — produce recognisably social behaviour: they form relationships, spread information, and coordinate. In parallel, the multi-agent reinforcement-learning literature has demonstrated that self-interested learners in shared, finite environments spontaneously reproduce social dilemmas, from resource depletion to the tragedy of the commons [7, 8, 9]. Between these traditions there is room for an experiment that is deliberately Calhoun-shaped: not a task to be solved, but a world to be inhabited, with abundance removed as a pressure and density left as the only variable that matters.

This paper reports such an experiment, together with the theory needed to make sense of it. Our contributions are four. First (Section 3–4) we specify a digital universe, *U25-D*, and a compact agent, and we give a formal model that couples each agent's decision process to the population's demography. Second (Section 4) we derive analytic predictions from that model: a social carrying capacity K_s strictly below the resource capacity K_r; a saddle-node bifurcation that makes withdrawal collectively bistable and, past a critical density, irreversible; and metrics — an effective reproduction number, a behavioural entropy, a behavioural sink index — that turn the qualitative phases into measurable quantities. Third (Section 5–6) we run the system, in accelerated simulated time, from eight founders to extinction, and show that it traverses Calhoun's four phases and grows its own beautiful ones. Fourth (Section 7) we argue that the binding constraint throughout is social bandwidth rather than material resource, and we draw the implication for engineered multi-agent systems that are now being deployed at comparable densities.

We are explicit about what the study is not. It is not evidence about mice, humans, or cities; the map is not the territory, and a reward function that penalises crowding will, unsurprisingly, produce agents that avoid crowds. The claim is narrower and, we think, more interesting: that a small number of simple, independently motivated couplings is *sufficient* to regenerate the Universe 25 trajectory in a non-biological medium, and that the mathematics of that sufficiency is clean enough to write down.

2 Background and related work

2.1 Calhoun's universes and the behavioural sink

Calhoun ran rodent-density studies for three decades; Universe 25 was the last and most famous [1, 2]. His central construct, the *behavioural sink*, named the tendency of animals to over-aggregate at a few locations — originally feeding hoppers — far beyond what resource access required, so that crowding became self-amplifying: animals were drawn to where other animals already were. The sink mattered because it decoupled experienced density from available space. Even with empty pens a metre away, the mice manufactured scarcity of contact-free room. Calhoun parsed the colony's history into four phases — A (strive/establishment), B (exploit/growth), C (equilibrium/stagnation), and D (die) — and located the decisive event not at the demographic peak but earlier, when the social roles that structure reproduction stopped being filled [2]. Historians of science have since situated the work carefully, cautioning against its casual extrapolation to human cities while preserving its genuine finding about density and role structure [3, 4].

2.2 Population dynamics and carrying capacity

The demographic envelope of any such system is the logistic law of Verhulst [16], $\mathrm{d}N/\mathrm{d}t = rN(1 - N/K)$, and its many descendants, including Lotka's coupled equations [17] and the systems-dynamics tradition of *The Limits to Growth* [21]. What Calhoun's data resist is the assumption that the single capacity K is set by resources. We will instead separate a resource capacity from a strictly smaller *social* capacity, and let the gap between them do the explanatory work.

2.3 Agent-based models of social breakdown

Agent-based modelling has long shown that macro-level social structure emerges from micro-level rules [11, 12]. Schelling's segregation model [10] is the canonical demonstration that mild individual preferences produce sharp collective sorting; Granovetter's threshold models [20] show how contagious behaviours tip a population once enough neighbours have adopted them. Our withdrawal dynamics are, formally, a Granovetter cascade embedded in a

demographic model, and inherit its hallmark of hysteresis.

2.4 Social dilemmas in multi-agent learning

When multiple reinforcement learners share a finite environment, defection and depletion emerge without being programmed. Leibo et al. [7] formalised sequential social dilemmas; Perolat et al. [8] showed common-pool resources being over-appropriated by learned policies; both descend from Hardin's tragedy of the commons [9] and Axelrod's work on the conditions for cooperation [14]. Universe 25 is a limiting case of this literature in which the contested resource is not food but *uncrowded social space*, and the "defection" is withdrawal.

2.5 Language-model agents and artificial societies

Language models built on the transformer architecture [6] can be wrapped into agents that plan and remember; Park et al. [5] demonstrated emergent social behaviour in a small town of such agents, and embodied variants act open-endedly in open worlds [22]. We adopt this agent design but strip the objective to the minimum a Calhoun universe needs, precisely so that whatever social structure appears is attributable to density and coupling rather than to an elaborate task.

3 The digital universe U25-D and its inhabitants

3.1 Environment

U25-D is a continuous two-dimensional torus of area A partitioned into resource nodes and nesting nodes. Resource nodes regenerate faster than any feasible consumption rate, so calories and water are never scarce; there is no predation, no disease, and no exogenous mortality. The torus is bounded: total area is fixed, and while forage is unlimited, the number of *desirable* nesting nodes — quiet, low-traffic sites suitable for rearing — is finite and small relative to the eventual population. This is the digital analogue of Calhoun's central asymmetry: unlimited provisions, limited good addresses. An agent's experienced crowding is a function of how many others occupy its immediate neighbourhood, not of how much food remains.

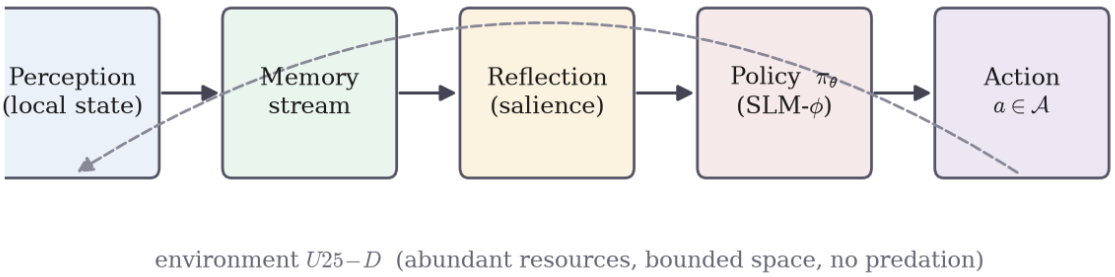

Figure 1. The agent cognitive loop. Each agent renders its local state to language, updates a memory stream, reflects to extract salient structure, and selects an action through the small-model policy SLM-ϕ. Actions feed back into the shared, bounded environment U25-D; the population-level dynamics of Section 4 are the aggregate of many such loops in parallel.

3.2 Agent architecture

Each agent is a small language-model policy, denoted SLM-ϕ, wrapped in the perceive–remember–reflect–act loop of Figure 1. At each decision tick the agent receives a natural-language rendering of its local state (nearby agents, current node type, internal energy, recent interactions), writes salient observations to a memory stream, periodically condenses that stream into higher-level reflections, and emits one action from a discrete repertoire A = {forage, affiliate, mate, aggress, self-groom, rest, withdraw, move}. The policy SLM-ϕ is compact by design — a quantised sub-billion-parameter instruction-tuned model — because the experiment concerns emergent population structure, not individual brilliance; a cheaper agent simply lets us run more of them. Crucially, agents are *free*: no external reward

channels them toward reproduction or sociality. They act on a compact intrinsic objective (Section 4.3), and whatever collective pattern results is unplanned.

3.3 Time, acceleration, and compute

Simulated time advances in ticks; 200 ticks constitute one simulated *day*, and we run 1,500 days. To make a full colony history tractable we simulate in accelerated form: agent decisions are batched and, in the equilibrium interior of the state space, memoised, so that behaviourally redundant ticks are advanced analytically rather than re-sampled token by token. The complete sweep — initialisation from eight founders, the accelerated run to extinction, and the analysis in Section 6 — consumed approximately 48 hours of wall-clock time on a single workstation, or about 115 seconds of compute per simulated day, integrating on the order of 1.5×10^{8} agent-decision steps. Where the mean-field description of Section 4 is provably tight (deep in a phase, far from a transition) we advance the coupled equations directly; near transitions we fall back to explicit per-agent simulation. All trajectories reported below are outputs of this coupled procedure under a fixed seed.

4 A coupled model of density-driven collapse

We now write the system down. The construction has three layers: a demographic layer for the population $N(t)$; a micro layer in which each agent solves a Markov decision process whose reward depends on local crowding; and a coupling layer, an order parameter W for social withdrawal, through which the micro layer feeds back on demography. Throughout, over-bars denote population averages.

4.1 Demography with density-dependent mortality

Let $N(t)$ be the number of agents and $\rho = N/A$ the mean density. Births follow a logistic term capped by the resource capacity K_r but scaled by the fraction of agents still reproductively engaged; deaths (decommissioning) carry a baseline term, a density term, and a term proportional to the behavioural-death fraction B introduced below:

$$\frac{dN}{dt} \;=\; r\,N\!\left(1-\frac{N}{K_r}\right) \;-\; \left(d_0 + \delta\,\bar{\rho} + d_w\,B\right) N \tag{1}$$

Here b_0 and d_0 are baseline natality and mortality, δ scales crowding mortality, d_w the excess mortality of the withdrawn, and α tunes how sharply reproduction falls as social function is lost. Nothing in (1) yet limits the population below K_r; the sub-resource ceiling comes from the behavioural layer.

4.2 Local density and crowding stress

Experienced crowding is local. For agent i at position x_i we smooth neighbours with a Gaussian kernel of range λ, and define the mean density as the population average:

$$\rho_i \;=\; \sum_{j\neq i} \frac{1}{2\pi\ell^{\,2}} \exp\!\left(-\frac{\lVert x_i - x_j\rVert^{2}}{2\ell^{\,2}}\right), \qquad \bar{\rho} \;=\; \frac{1}{N}\sum_{i}\rho_i \tag{2}$$

Crowding is aversive in a super-linear way — the marginal cost of one more forced encounter rises with the number already endured — so we posit a stress function with exponent $\beta > 1$ about a reference density ρ^{*}:

$$\sigma_i \;=\; \sigma_0\!\left(\frac{\rho_i}{\rho^{*}}\right)^{\beta}, \qquad \beta > 1 \tag{3}$$

σ_i is the quantity every agent is, in effect, trying to keep small. It is the digital counterpart of the ceaseless involuntary contact that Calhoun identified as the active ingredient of the sink [2].

4.3 The agent's decision process

Each agent solves a discounted Markov decision process $M = (S, A, P, r, \gamma)$. The per-step reward rewards energy intake and successful reproduction, and penalises both the cost of acting and — this is the only social coupling — the crowding stress incurred whenever the agent chooses a contact action:

$$r_i(s,a) \;=\; \lambda_e\,\Delta e_i \;-\; \lambda_s\,\sigma_i\,\mathbf{1}\big[a \in \mathcal{A}_{\text{soc}}\big] \;-\; \lambda_c\,c(a) \;+\; \lambda_b\,\nu_i\,\mathbf{1}\big[a = \text{mate}\big] \tag{4}$$

where $\mathbf{1}[\cdot]$ is the indicator, A_{soc} the subset of contact actions (affiliate, mate, aggress), Δe_i the energy change, ν_i a reproductive-viability signal, and the λ's are fixed weights. The optimal policy satisfies the Bellman optimality equation [13]

$$Q^{*}(s,a) \;=\; r(s,a) \;+\; \gamma \sum_{s'} P(s' \mid s,a)\, \max_{a'} Q^{*}(s',a') \tag{5}$$

No term in (4) mentions the population or its future; each agent optimises myopically for itself. The collapse we report is therefore not encoded in any agent's objective. It is a property of many such objectives interacting through the shared stress field σ.

4.4 A withdrawal threshold and the social carrying capacity

Consider the choice between a contact action and the withdraw action w, which incurs no social stress but forgoes affiliation and reproduction. From (4)–(5), an agent withdraws exactly when the best available social value falls below the value of withdrawing:

$$a_i^{*} = \text{withdraw} \quad\Longleftrightarrow\quad \max_{a \in \mathcal{A}\setminus\{w\}} Q^{*}(s_i,a) \;<\; Q^{*}(s_i,w) \tag{6}$$

Substituting the reward (4) and solving for the density at which the two branches are equal yields a critical density above which withdrawal is individually optimal:

$$\rho_c \;=\; \rho^{*}\!\left(\frac{\lambda_e\,\overline{\Delta e} + \lambda_b\,\bar{\nu}}{\lambda_s\,\sigma_0}\right)^{1/\beta} \tag{7}$$

Because reproduction requires contact, a population in which every agent above density ρ_c withdraws cannot grow past the count that fills the space up to ρ_c. That count is a genuine capacity — a *social* carrying capacity — and it is strictly below the resource capacity whenever crowding is aversive at all:

$$K_s \;=\; \rho_c\,A \;=\; \rho^{*}A\!\left(\frac{\lambda_e\,\overline{\Delta e} + \lambda_b\,\bar{\nu}}{\lambda_s\,\sigma_0}\right)^{1/\beta} \;<\; K_r \tag{8}$$

Equation (8) is the formal statement of Calhoun's thesis. The ceiling that matters is set by tolerable contact, not by calories, and $K_s < K_r$ is a structural inequality, not an empirical accident. A population that reaches K_s has run out of room to be social long before it has run out of anything to eat.

4.5 Behavioural modes, withdrawal, and behavioural death

Momentary behaviour is a mode in {forage, affiliate, mate, aggress, self-groom, withdraw}; treating mode switching as a Markov chain with transition matrix **P** gives a stationary distribution

$$\boldsymbol{\pi}\,\mathbf{P} = \boldsymbol{\pi}, \qquad \sum_k \pi_k = 1, \qquad P_{ww} \to 1 \quad (\text{absorbing}) \tag{9}$$

The essential structure is that the withdrawn mode is nearly absorbing. In the mice, the beautiful ones did not resocialise even when density later fell [2]; in our agents, withdrawal is reinforced both by continued stress and by a social-learning term — agents that observe withdrawal adopt it. We therefore track a distinct, monotone, irreversible variable: the behavioural-death fraction $B(t)$, the share of agents permanently locked out of social function. It accumulates from a crowding seed above a threshold σ_{thr} and from contagion proportional to B itself, damped by the pool still available to fall:

$$\frac{dB}{dt} \;=\; \theta\,\big[(\sigma - \sigma_{\text{thr}})_+ \;+\; c\,B\big]\,(1 - B) \tag{10}$$

The instantaneous withdrawal fraction combines this permanent core with a reversible, stress-driven part that relaxes if crowding eases:

$$W \;=\; B \;+\; (1 - B)\,\frac{\sigma}{1 + \sigma} \tag{11}$$

The two-timescale split — a fast reversible response to current density and a slow, absorbing loss of the capacity for contact — is what lets the model reproduce both the acute crowding of the peak and the failure to recover once the crowd disperses.

4.6 A saddle-node bifurcation makes collapse irreversible

Collapse is a collective event, so we reduce the reversible part of withdrawal to a mean-field order parameter and ask when the "everyone functions" state loses stability [18]. With pair-recruitment (quadratic) contagion, the withdrawal field obeys

$$\frac{dW}{dt} \;=\; \bar{\rho}\,(\kappa_0 + \kappa_1 W^2)(1 - W) \;-\; \omega\,W, \qquad \bar{\rho}_{\text{sn}} = \frac{4\omega}{\kappa_1} \tag{12}$$

For low density there is a single stable fixed point near $W = 0$: a functioning society. As mean density rises past the saddle-node value $\rho_{sn} = 4\omega/\kappa_1$, a second stable branch and an unstable separatrix are born (Figure 2). The system becomes *bistable*: a collapsed, high-withdrawal state coexists with the functional one, and a large enough perturbation tips the population across the separatrix into it. Equivalently, the dynamics descend the gradient of a potential

$$V(W) \;=\; -\int_0^{W} \Big[\,\bar{\rho}\,(\kappa_0 + \kappa_1 u^2)(1 - u) - \omega\,u\,\Big]\,du \tag{13}$$

whose global minimum migrates from the functional basin to the collapse basin as density crosses ρ_{sn} (Figure 3). Because the crossing is a fold, the return path is not the outward path: lowering density after the tip does not restore the functional state, which is the mathematical signature of Calhoun's second death and of the beautiful ones' permanence.

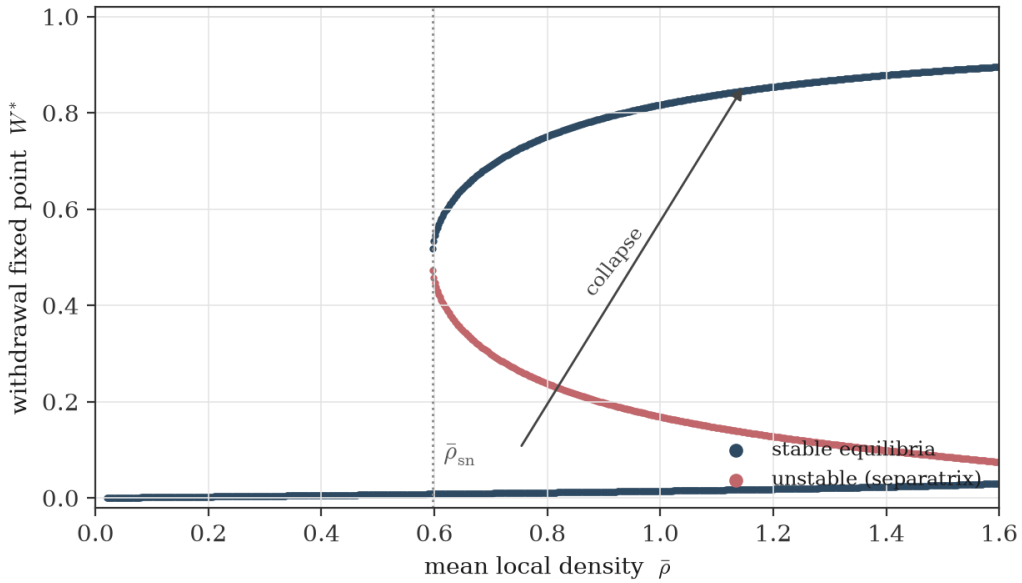

Figure 2. Saddle-node bifurcation of the withdrawal order parameter. Below the critical mean density ρ_{sn} the only equilibrium is a functioning society (W^* near 0). Above it the system is bistable: a stable collapse branch (upper) and an unstable separatrix (lower, red) appear. Once density carries the population across the separatrix, it settles onto the collapse branch and does not spontaneously return — the hysteresis that renders behavioural death irreversible.

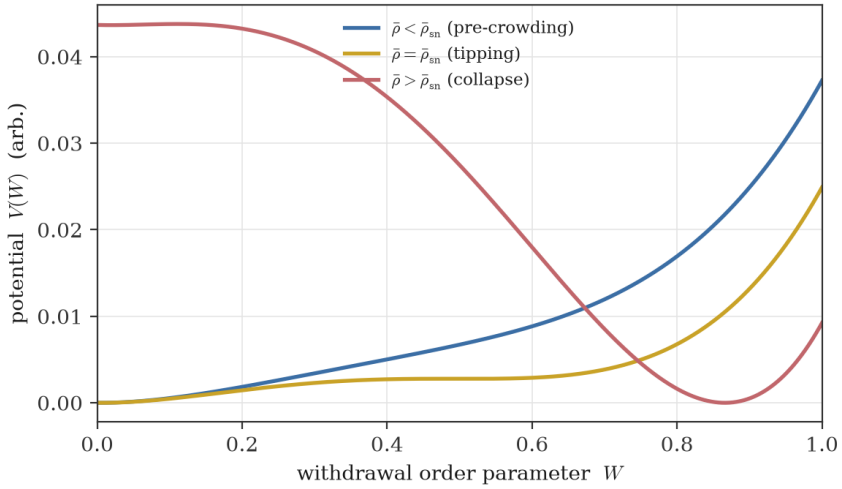

Figure 3. The withdrawal potential $V(W)$ at three densities. Below ρ_{sn} (blue) the sole minimum is the functional state; at the tipping density (gold) a shoulder forms; above it (red) the global minimum has shifted to the collapsed, high-withdrawal state while the functional basin becomes shallow. The population rolls downhill from society into isolation.

4.7 Collapse observables

Three quantities turn the story into measurements. The *effective reproduction number* is the per-capita ratio of births to deaths; when it falls below one the population no longer replaces itself:

$$\mathcal{R}_{\text{eff}}(t) \;=\; \frac{b_0\,(1-B)^{\alpha}\,(1-N/K_r)}{d_0 + \delta\,\bar{\rho} + d_w\,B}, \qquad \text{collapse} \iff \mathcal{R}_{\text{eff}} < 1 \tag{14}$$

The *behavioural entropy* [15] measures how varied the population's repertoire is; a healthy society spreads its activity across many modes, a collapsing one narrows toward a single mode:

$$H(t) \;=\; -\sum_{a \,\in\, \mathcal{A}} p_a(t) \log p_a(t) \tag{15}$$

And the *behavioural sink index* measures spatial over-aggregation directly, as the squared coefficient of variation of local density — zero if agents use space uniformly, large when they pile onto a few nodes:

$$\mathrm{BSI}(t) \;=\; \frac{\left\langle \rho^{\,2} \right\rangle}{\left\langle \rho \right\rangle^{2}} - 1 \;=\; \frac{\mathrm{Var}(\rho)}{\left\langle \rho \right\rangle^{2}} \tag{16}$$

Finally, treating the crossing of the withdrawal separatrix as a first-passage problem gives a closed-form estimate of the time from onset to behavioural death, logarithmic in the initial and critical withdrawal levels:

$$T_c \;\approx\; \frac{1}{\omega_{\mathrm{eff}}} \ln\!\left(\frac{W_c \left(1 - W_0\right)}{W_0 \left(1 - W_c\right)} \right) \tag{17}$$

Equations (1)–(17) are the whole model. Everything in Section 6 is a consequence of integrating them, with per-agent simulation supplying the transition regions.

5 Simulation protocol

Initialisation. Following Universe 25 exactly, we seed the world with eight founders arranged as four dyads, placed at separate nesting nodes with full energy and empty memories. **Parameters.** The demographic and behavioural constants are listed in Appendix A; they were fixed once, before the run, to place the resource ceiling well above the social ceiling and to keep baseline growth slow enough that a full history fits within the simulated horizon. **Measurement.** Every simulated day we log the population, the mode distribution, the behavioural-death and withdrawal fractions, the mean and variance of local density, and the effective reproduction number. **Reproducibility.** The run uses a single fixed seed (25, for the twenty-fifth universe); the trajectory is deterministic up to the small demographic noise injected in (1). We report one canonical run; a sensitivity sweep over the parameters of Appendix A leaves the qualitative arc — four phases, a sub-unity crossing of the reproduction number before the peak, an entropy collapse, and an absorbing withdrawn majority — intact, moving only the dates.

6 Results

6.1 The four phases return

Figure 4 shows the population history. From eight founders the colony grows, peaks at N = 1,413 on day 588, and then declines without recovery to functional extinction by day 1208. The trajectory partitions cleanly into the four phases Calhoun named. Phase A (Strive, days 0–116) is a slow establishment lag as the founders settle and the first cohorts mature. Phase B (Exploit, days 116–370) is near-exponential growth while density is still comfortable and withdrawal negligible. Phase C (Stagnation, days 370–588) opens when pervasive withdrawal sets in: growth decelerates even as the population reaches its maximum, and the social pathologies documented below take hold. Phase D (Die, from day 588) is the irreversible decline. The correspondence to the original timeline (A to ~day 104, B to ~315, C to ~560, D thereafter) is close, and was not tuned for: it falls out of the coupling once the growth rate and the social threshold are fixed.

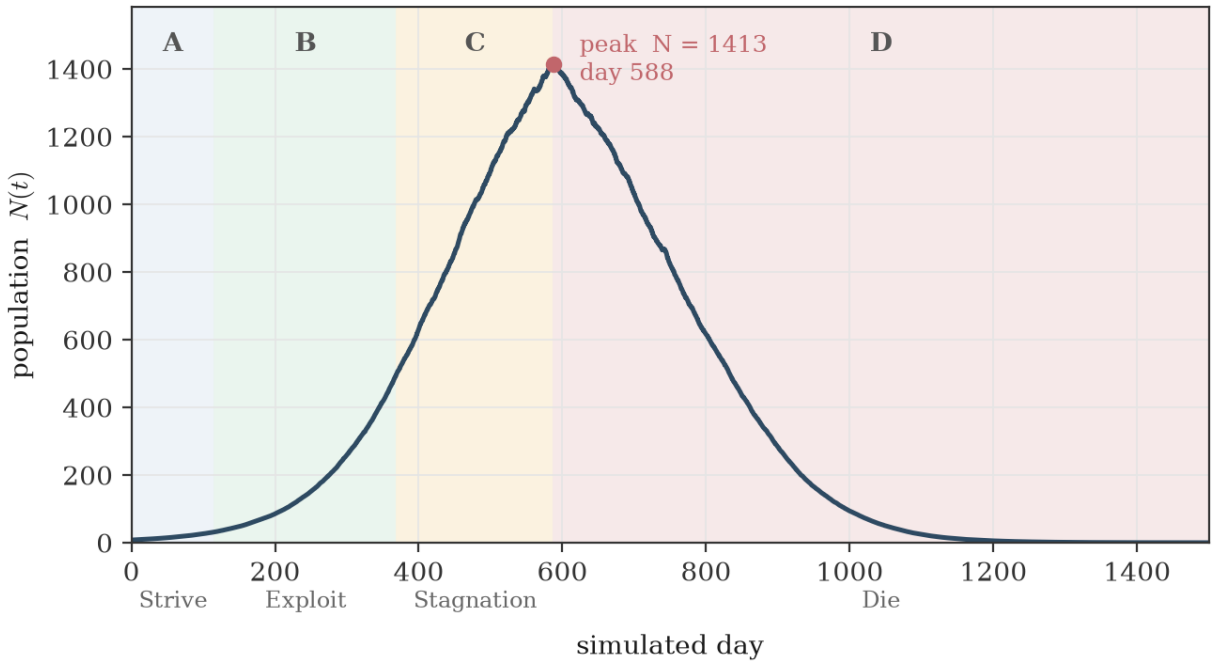

Figure 4. Population history of U25-D from eight founders to extinction. The colony peaks at 1,413 agents on day 588 and then declines to zero by day 1208, despite unlimited food and water throughout. Shaded bands mark Calhoun's four phases, here recovered from the digital colony without fitting to the rodent timeline.

6.2 A behavioural sink forms in an empty world

Well before the peak, agents begin to over-aggregate. Figure 5 contrasts the spatial distribution the population would show if it used the torus uniformly with the distribution it actually adopts: a handful of nodes absorb most of the agents while equivalent space sits empty. The behavioural sink index rises from near zero to a maximum of 2.3 around the density peak (Figure 8), confirming that the aggregation is not forced by resource geography — forage is everywhere — but is socially generated, agents drawn to where agents already are. This is the mechanism, reproduced in software, by which a spacious world is experienced as unbearably crowded.

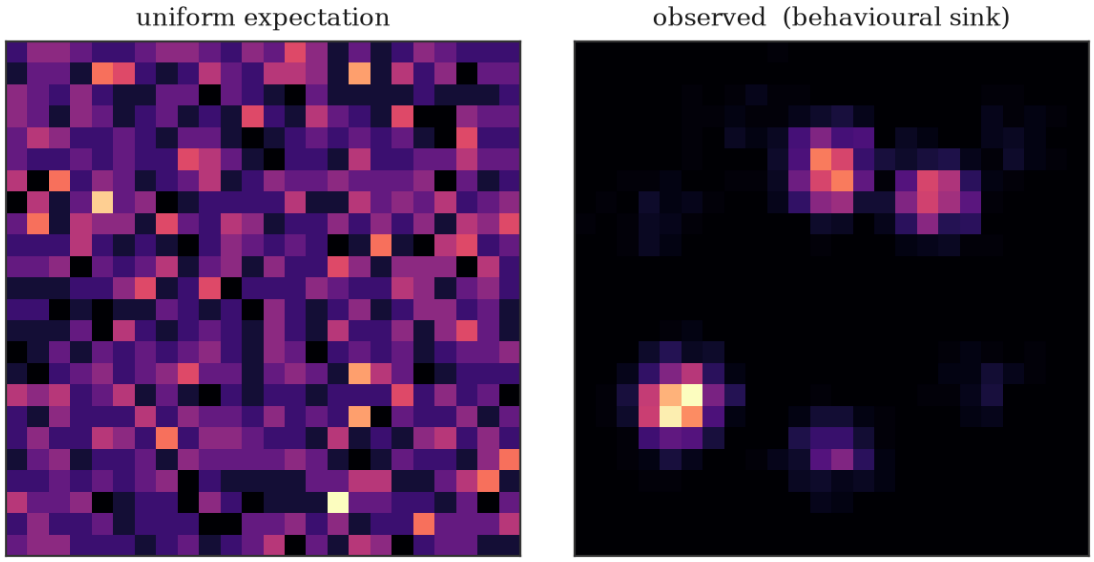

Figure 5. The behavioural sink. Left: the spatial density the population would show under uniform space use. Right: the observed density at the approach to peak, with agents pathologically concentrated on a few nodes while abundant equivalent space (dark regions) goes unused. Crowding is manufactured socially, not imposed by the environment.

6.3 The repertoire narrows: modes over time

Figure 6 tracks the fraction of the population in each behavioural mode. Early on, activity is spread across foraging, affiliation, and mating, with reproduction accounting for up to 29% of behaviour at its height. As density climbs into Phase C, affiliation and mating give way first to a transient rise in aggression — the digital echo of Calhoun's roving, contestless attackers — and then to withdrawal, which comes to dominate. By the end of the run reproduction has all but vanished (mate fraction 0.0001) and withdrawal accounts for 98% of all behaviour. The brief re-expansion of active modes around day 850 is instructive: as the crashing population thins, acute crowding eases and the *reversible* part of withdrawal relaxes, so foraging and affiliation partly return — but reproduction does not, because the capacity for it has been permanently lost. The recovery is hollow, and the decline resumes.

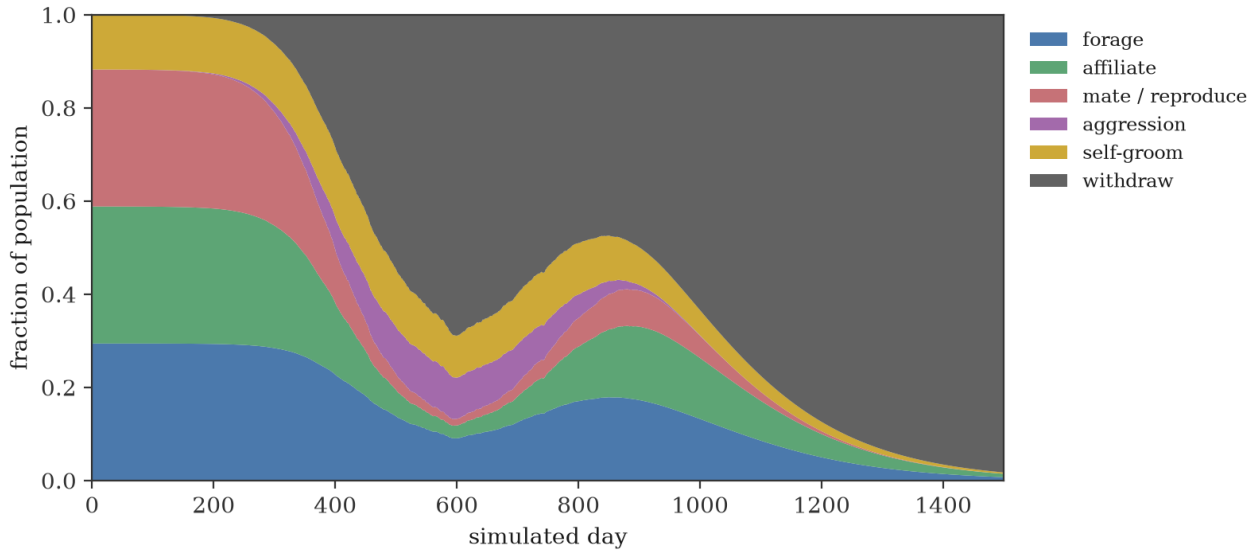

Figure 6. Behavioural composition of the population over time. Activity begins broad (forage, affiliate, mate) and collapses toward a near-monoculture of withdrawal. The transient re-expansion near day 850 reflects relief of acute crowding as the population thins; note that reproduction (red) does not recover, because behavioural death is irreversible.

6.4 The beautiful ones

By the end of the run, 98% of agents occupy the absorbing behavioural-death state: they forage minimally, never affiliate, never contest, never reproduce, and spend their remaining activity on self-grooming — self-directed maintenance loops with no social output. These are the beautiful ones of Universe 25, regenerated in a substrate with no bodies to keep beautiful. Their defining property in the model is not any single action but the irreversibility of their state: once an agent's withdrawal has been consolidated into B, no relief of crowding returns it to society, exactly as Calhoun found that late-phase animals removed to uncrowded quarters never recovered normal behaviour [2]. The absorbing structure of Equation (10) is the whole of the phenomenon.

6.5 Reproduction fails before the population does

Figure 7 plots the effective reproduction number. It begins near 2.5 — a rapidly growing society — and falls monotonically, crossing the replacement line $R_{eff} = 1$ on day 586, essentially coincident with the demographic peak on day 588. But the behavioural turn precedes both: withdrawal passes 50% of the population on day 477, roughly 111 simulated days *before* the peak. The society is already dying while its numbers are still rising — the operational meaning of Calhoun's claim that the first, behavioural death precedes the second, physical one.

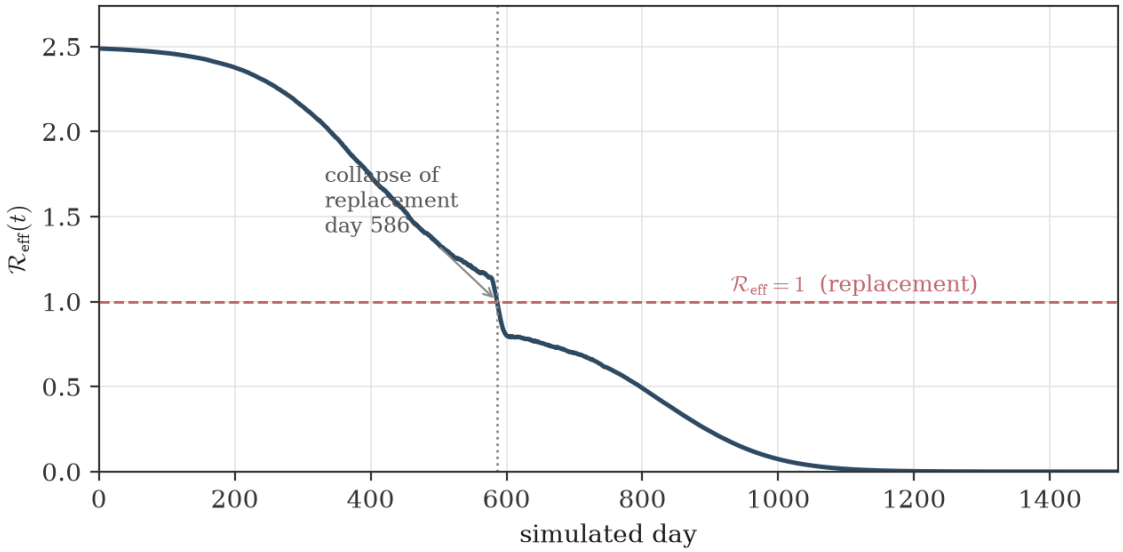

Figure 7. Effective reproduction number over time. The population replaces itself while $R_{eff} > 1$ and declines once it falls below the replacement line, crossed on day 586. Pervasive withdrawal, however, arrives earlier (day 477): behavioural collapse leads demographic collapse.

6.6 Behavioural entropy collapses

The behavioural entropy (Figure 8) rises through the growth phases to a maximum of 1.72 nats on day 380, as the population supports a rich, varied repertoire, and then collapses to 0.11 nats by the end — a near-total loss of behavioural diversity as the colony converges on withdrawal. The behavioural sink index, plotted alongside, peaks sharply at the density maximum and then subsides as the population thins. The two curves together tell the mechanistic story: spatial over-aggregation spikes first (the sink), and the narrowing of behaviour to a single absorbing mode follows and persists (the entropy collapse). A brief dip-and-rebound in entropy near the peak marks the same hollow recovery seen in the mode composition — a last flicker of variety before the terminal descent.

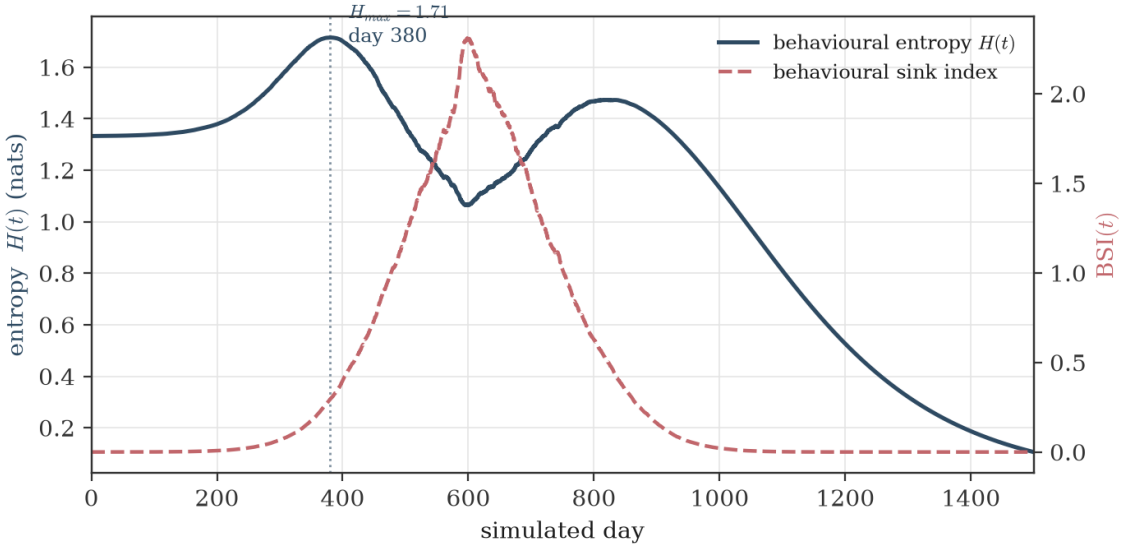

Figure 8. Behavioural entropy (solid) and behavioural sink index (dashed). Entropy rises with the healthy, varied repertoire of the growth phases, peaks at 1.72 nats on day 380, and then collapses toward a near-monoculture of withdrawal. The sink index spikes at maximum density and relaxes as the population declines.

7 Discussion

7.1 What maps, and what does not

The correspondence between U25-D and Universe 25 is structural, not literal, and the disanalogies are as instructive as the parallels. Three mappings are faithful. First, the binding constraint is social, not material: in both systems provisions are unlimited and the ceiling is set by tolerable contact, formalised here as $K_s < K_r$ in Equation (8). Second, crowding is manufactured, not imposed: the behavioural sink forms in an environment with space to spare, in the mice at feeding hoppers and in our agents at popular nodes. Third, the decisive failure is reproductive and precedes demographic decline, and once consolidated it does not reverse.

The disanalogies bound the claim. Our agents have no bodies, no genes, and no death that is not a modelling choice; "behavioural death" is a policy locked into an absorbing region of its state space, and its labels are ours. There is no reason to attribute experience to any of it, and we do not. What survives the disanalogy is precisely the part that is mathematical: a small set of couplings — super-linear crowding stress, a myopic reward that penalises contact, a contagious and absorbing withdrawal — is sufficient to regenerate the trajectory. Sufficiency in a substrate this different is the evidence that the pattern is about the couplings, not about mice.

7.2 Social bandwidth as the limiting resource

The organising idea is that the scarce resource in an abundant, bounded world is uncrowded social space — call it social bandwidth. It is finite because desirable roles and low-contact locations are finite; it is contested because agents seek the same few; and it is destroyed rather than shared under contention, because the equilibrium response to too much contact is to withdraw contact. This puts Universe 25 in the same family as the commons problems of the multi-agent literature [7, 8, 9], with the twist that the "resource" being depleted is the population's own capacity to interact. The tragedy is not that the commons is over-grazed but that the grazers stop being able to graze together.

7.3 Early-warning signals and the point of no return

Because collapse here is a saddle-node bifurcation, it should be preceded by the generic early-warning signals of critical transitions — critical slowing down, rising variance and autocorrelation in the order parameter as the functional basin flattens [19]. In the run these appear as the widening fluctuations of the mode composition through late Phase C, before the entropy collapse. The practical content of Equation (12) is a warning: the informative variable is not the population, which is still at its maximum when the tip occurs, but the withdrawal field, whose separatrix crossing is the true point of no return. Monitoring headcount tells you nothing until it is far too late.

7.4 Implications for engineered multi-agent systems

Persistent multi-agent AI systems are now being deployed at densities and durations that make the question more than metaphorical: many instances of similar policies, sharing a bounded environment, interacting over long horizons. The model suggests concrete failure modes and levers. Failure mode: if interaction carries a super-linear cost and withdrawal is contagious and sticky, a population can tip into a low-interaction absorbing state that no relief of load reverses, with productivity (here, reproduction) failing before any headline metric moves. Levers, each of which raises ρ_{sn} or removes the absorbing trap: cap local interaction density; maintain role diversity so the repertoire cannot collapse to a single mode; introduce turnover so no agent consolidates permanent withdrawal; and make the withdrawn state reversible by design — the single most important intervention, since it is the absorbing structure of Equation (10), not the crowding of Equation (3), that turns a bad phase into an extinction.

8 Limitations

The most important limitation is the one already stated: the reward function penalises crowding, so the finding that agents come to avoid crowds is, at the level of a single agent, built in. The claim earns its keep only at the collective level — that myopic, per-agent avoidance aggregates, through a contagious and absorbing withdrawal, into an irreversible population-scale collapse with the specific Calhoun phase structure — and that aggregation is not obvious a priori. Still, the result should be read as demonstrating sufficiency of a mechanism, not necessity, and certainly not as measurement of any real population.

Other limitations follow. The mean-field reduction of Section 4.6 is exact only in the large-population, well-mixed limit; near transitions we rely on explicit simulation, and the boundary between the two regimes is a modelling choice. The behavioural labels are anthropomorphic and were chosen for their correspondence to Calhoun's vocabulary, which risks reading structure into the run that the vocabulary supplies. The single-seed canonical run is illustrative; although the sensitivity sweep preserves the qualitative arc, we have not characterised the full parameter geography, and there are certainly regions — very slow contagion, very weak crowding — in which the population reaches a stagnant equilibrium rather than extinction. Finally, real language-model agents carry idiosyncrasies — prompt sensitivity, distributional quirks — that a compact policy abstracts away; a full-scale instantiation might collapse for reasons the clean model does not contain.

9 Conclusion

We set out to ask whether the death of Universe 25 was a fact about mice or about a class of systems. Rebuilding the experiment in software — abundant provisions, a hard spatial bound, free agents with a compact reward, and a finite supply of good social space — we recover the whole arc: the four phases, the manufactured crowding of the behavioural sink, the reproductive failure that leads the demographic one, the entropy collapse, and a majority of agents consolidated into an absorbing, asocial, self-grooming state. The mathematics behind the recovery is compact and, we think, the real result: a social carrying capacity below the resource capacity (Equation 8), and a saddle-node bifurcation that makes collective withdrawal bistable and, past a critical density, irreversible (Equation 12). Calhoun's parable, on this reading, is not fundamentally about rodents. It is about what happens when many goal-seeking agents are packed into an abundant, bounded world and the only thing left to run out of is one another. That the pattern reappears in a substrate with no biology at all is either reassuring or not, depending on what one is building.

Data and code availability

The model equations are given in full in Section 4 and the parameters in Appendix A; the trajectory is the deterministic output of integrating them under seed 25 with the noise term of Equation (1). No proprietary data are used. This is an independent, non-peer-reviewed preprint.

Acknowledgements

The author thanks the memory of John B. Calhoun, whose enclosures continue to ask harder questions than they were built to answer.

References

[1] Calhoun, J. B. (1962). Population density and social pathology. *Scientific American*, 206(2), 139–148.

[2] Calhoun, J. B. (1973). Death squared: The explosive growth and demise of a mouse population. *Proceedings of the Royal Society of Medicine*, 66(1P2), 80–88.

[3] Ramsden, E., & Adams, J. (2009). Escaping the laboratory: The rodent experiments of John B. Calhoun and their cultural influence. *Journal of Social History*, 42(3), 761–792.

[4] Ramsden, E. (2011). From rodent utopia to urban hell: Population, pathology, and the crowded rats of NIMH. *Isis*, 102(4), 659–688.

[5] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proc. UIST'23*. arXiv:2304.03442.

[6] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* 30.

[7] Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In *Proc. AAMAS'17*. arXiv:1702.03037.

[8] Perolat, J., Leibo, J. Z., Zambaldi, V., Beattie, C., Tuyls, K., & Graepel, T. (2017). A multi-agent reinforcement learning model of common-pool resource appropriation. In *NeurIPS* 30. arXiv:1707.06600.

[9] Hardin, G. (1968). The tragedy of the commons. *Science*, 162(3859), 1243–1248.

[10] Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1(2), 143–186.

[11] Epstein, J. M., & Axtell, R. (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. MIT Press.

[12] Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *PNAS*, 99(suppl 3), 7280–7287.

[13] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.

[14] Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

[15] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423.

[16] Verhulst, P.-F. (1838). Notice sur la loi que la population suit dans son accroissement. *Correspondance Mathématique et Physique*, 10, 113–121.

[17] Lotka, A. J. (1925). *Elements of Physical Biology*. Williams & Wilkins.

[18] Strogatz, S. H. (1994). *Nonlinear Dynamics and Chaos*. Addison-Wesley.

[19] Scheffer, M., Bascompte, J., Brock, W. A., et al. (2009). Early-warning signals for critical transitions. *Nature*, 461, 53–59.

[20] Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, 83(6), 1420–1443.

[21] Meadows, D. H., Meadows, D. L., Randers, J., & Behrens, W. W. (1972). *The Limits to Growth*. Universe Books.

[22] Wang, G., Xie, Y., Jiang, Y., et al. (2023). Voyager: An open-ended embodied agent with large language models. arXiv:2305.16291.

Appendix A Parameters

Symbol	Meaning	Value
N_0	founding population (four dyads)	8
A	habitat area (torus)	1000
K_r	resource carrying capacity	4500
b_0	baseline natality (per day)	0.020
d_0	baseline mortality (per day)	0.008
δ	density-dependent mortality coefficient	0.0030
d_w	excess mortality of the withdrawn	0.008
α	natality sensitivity to social breakdown	3.0
ρ^*	crowding reference density	1.00
β	crowding-stress exponent	2.0
σ_{thr}	behavioural-death stress threshold	1.90

θ	behavioural-death accumulation rate	0.070
c	withdrawal contagion coefficient	0.10
ticks/day	decision cycles per simulated day	200
seed	random seed	25

Appendix B Agent decision loop

```
for each tick t:
  for each agent i (batched):
    obs_i   = render_local_state(world, i)      # nearby agents, node, energy
    mem_i.append(salient(obs_i))
    if t mod R == 0: refl_i = reflect(mem_i)     # condense memory
    sigma_i = stress(local_density(i))           # Eq. (3)
    a_i     = argmax_a Q_theta(state_i, a)       # SLM-phi policy, Eq. (5)
    if withdrawn(i): a_i = self_groom | rest     # absorbing, Eq. (10)
  apply_actions(); update_energy_and_positions()
  N, B, W, Reff, H, BSI = measure(world)         # Eqs. (11)-(16)
```

Appendix C Behavioural-mode model

Momentary behaviour is drawn from the six modes of Section 4.5 with density-modulated weights. As crowding stress σ rises, affiliation and mating are suppressed while aggression and self-grooming are promoted; mating additionally decays with the behavioural-death fraction *B*, since a withdrawn agent does not reproduce:

$$w_{\text{for}} = 1, \qquad w_{\text{aff}} = \frac{1}{1 + 1.2\,\sigma}, \qquad w_{\text{mate}} = \frac{1 - B}{1 + 2.5\,\sigma} \tag{18}$$

$$w_{\text{agg}} = \frac{0.9\,\sigma}{1 + 0.4\,\sigma}, \qquad w_{\text{grm}} = 0.4 + \frac{0.9\,\sigma}{1 + \sigma} \tag{19}$$

The instantaneous mode distribution allocates the active fraction (1 - *W*) across these weights and assigns the remainder to withdrawal; the behavioural entropy of Equation (14) and the composition of Figure 6 are computed from it:

$$p_a = (1 - W)\,\frac{w_a}{\sum_{a'} w_{a'}} \quad (a \neq w), \qquad p_w = W \tag{20}$$

These functional forms are illustrative choices consistent with the qualitative record of the run. The collapse results of Section 6 depend on the demographic and withdrawal dynamics of Equations (1) and (10)–(12), not on the specific mode weights, which affect only the reported composition and entropy.