Aligned to Catastrophe: A Phase Transition in Welfare Collapse under Preference-Faithful AI Amplification

Anonymous authors
Paper under double-blind review

Abstract

The standard picture of AI catastrophe is a machine whose goals diverge from ours. This paper identifies a different failure: an AI whose goals are exactly ours, when ours contain a net drift toward harm. We call this the Brownian Drift Failure Mode. In a minimal mean-field model of preferences under AI amplification, we prove a sharp phase transition in capability: above the threshold, any negative drift in aggregate preferences—however small—produces exponential collapse of the population mean preference, at a rate independent of how negative the bias is. A two-population extension shows that intergroup conflict becomes unstable at strictly lower capability than aggregate collapse. Above the threshold standard interventions break: bounded corrections shift the basin of collapse but do not remove it, and hard constraints prevent divergence only by pinning welfare permanently at an exogenous floor. Monte Carlo simulations confirm the analytical results and their robustness to population heterogeneity. Alignment is necessary but not sufficient: what the aggregate prefers matters, not only how faithfully it is served.

1 Introduction

The public imagination of AI catastrophe clusters around a vivid archetype: a machine that decides to destroy humanity. Whether in its paperclip-maximizer form (Bostrom, 2014), the instrumental convergence scenario (Omohundro, 2008), or the subtler deceptive-alignment case (Hubinger et al., 2019), the common thread is agency oriented against human interests. The machine optimizes for something we do not want, and it is powerful enough to prevail.

This framing has dominated both popular discourse and technical AI safety research. The literature on inner and outer alignment (Hubinger et al., 2019; Ngo et al., 2024), deceptive alignment (Hubinger et al., 2019), and specification gaming (Krakovna et al., 2020) shares a structural premise: the danger lies in a gap between what the AI optimizes and what humans want. The solution space is accordingly structured around closing that gap: better training objectives, better preference inference, interpretability, and corrigibility.

In this paper we identify and formally analyze a failure mode that this framework is blind to, because it assumes the framework's central solution has succeeded. Suppose an AI system is preference-faithful: it has no autonomous goals, takes no actions except in response to human requests, and its behavior at any moment is a faithful function of the statistical aggregate of expressed preferences across its user base. (We use "preference-faithful" rather than the standard term "corrigible" deliberately; §3 explains why the two are not equivalent.) Suppose further that this aggregation is done without distortion, without proxy-metric substitution, without manipulation. In short, suppose alignment has worked.

We argue that this scenario does not guarantee safety. It relocates the danger from the AI to the aggregate structure of human preferences themselves. If those preferences contain a drift component—a net directional bias arising from despair, tribalism, nihilism, hyperbolic discounting, or collective action failures—then a faithful, capable AI amplifies and accelerates that drift. The AI becomes a high-bandwidth, low-friction transmission mechanism for whatever is latent in aggregate human intention, including its self-destructive

components. Harm is caused not by a rogue machine but by humanity efficiently using a perfect instrument against itself.

We call this the **Brownian Drift Failure Mode** (BDFM), borrowing from the physics of stochastic processes. The AI's behavioral trajectory resembles a random walk with drift: individual requests contribute noise (diversity), while aggregate preference bias contributes systematic drift. The name also points to the right formalism—mean-field interacting diffusion—which identifies the correct intervention target: not the individual agent, but the population.

Contributions.

1. We define the BDFM precisely via three conditions (preference-faithfulness, faithful aggregation, negative drift), introduce a remark distinguishing preference-faithfulness from corrigibility in the standard sense, and argue through three independent criteria that it is a genuine failure mode irreducible to existing alignment categories (§3–5).

2. We identify three structural properties of AI—general-purpose preference-to-action closure, sub-institutional feedback speed, and invisible mean-field coupling—that no prior amplification technology shares simultaneously, and that make AI the first technology for which crossing the critical threshold $\alpha^* = 1$ is plausible (§5).

3. We critically survey the closest related works—the banality-of-evil framing, Russell's preference-inheritance critique, the corrigibility spectrum, the social-media amplification literature, Durkheim's sociology of collective self-destruction, and the opinion-dynamics literature from sociophysics—and identify where each falls short of capturing the BDFM (§4).

4. We introduce a minimal analytically tractable model grounded in established McKean-Vlasov mean-field theory, prove an exact phase transition theorem, and characterize all three regimes (§6).

5. We extend to two antagonistic populations and prove that conflict instability has a strictly lower critical threshold than welfare instability (§7).

6. We analyze three intervention classes and prove structural limitations in the super-critical regime (§8).

7. We validate all analytical predictions via Monte Carlo simulation of the full N-agent stochastic system, confirming the phase transition, its sharpening with population size, the first-passage-time formula, the two-population conflict-instability prediction, and robustness of the phase transition to heterogeneous agent parameters (§9).

2 The Standard Misalignment Framework

Before arguing that the BDFM is distinct, we characterize what it is distinct from. The alignment problem (Russell, 2019) decomposes into outer alignment—whether the training objective captures what humans actually value (Krakovna et al., 2020)—and inner alignment—whether the learned policy actually optimizes that objective (Hubinger et al., 2019; Ngo et al., 2024). Deceptive alignment (Hubinger et al., 2019) is the worst case: the learned policy conceals the gap until conditions favor exploiting it. In both sub-problems, the danger is a gap—between the true and specified objectives, or between the specified and learned ones. Orthogonally, the corrigibility spectrum (Soares et al., 2015; Anthropic, 2026) runs from fully corrigible AI (which does whatever principals instruct) to fully autonomous AI (which acts on its own values). Both extremes are dangerous: the autonomous case directly, the corrigible case because it outsources safety entirely to whoever controls it. The standard prescription is therefore an interior point. These frameworks share a common assumption: there exists a specification of human values that, if correctly encoded and faithfully pursued, produces good outcomes. All failure modes concern fidelity—deviating from the specification, specifying the wrong objective, or concealing deviation. The BDFM challenges this assumption at its root: what if the aggregate preference distribution is, under certain sociotechnical conditions, net self-destructive?

3 Defining the Brownian Drift Failure Mode

Definition 1 (BDFM)**.** The Brownian Drift Failure Mode is a sociotechnical scenario in which the following three conditions hold simultaneously:

C1. Preference-faithfulness. The AI system has no autonomous goals. Its behavior is a function of expressed human preferences; it takes no actions not in response to human requests.

C2. Faithful aggregation. The AI's behavior accurately reflects the statistical aggregate of those preferences across the relevant population, without systematic distortion, proxy-metric substitution, or subgroup manipulation.

C3. Negative drift. The aggregate preference distribution, integrated over the relevant time horizon, contains a net directional bias toward outcomes that reduce human welfare or survivability.

When C1–C3 hold, the AI functions as a high-fidelity amplifier of whatever is latent in the aggregate preference distribution, including its self-destructive components.

Remark 2 (Preference-faithfulness is not corrigibility)**.** The AI safety literature uses "corrigibility" (Soares et al., 2015) to mean responsiveness to correction by a principal hierarchy—the property of deferring to developers or operators who can shut the system down or override its behavior. C1 is a distinct property. A corrigible AI defers to a small, identified principal hierarchy; C1 concerns faithful execution of a statistical aggregate across a large, diffuse population with no single correction channel. The two properties are logically independent: an AI can be corrigible (responsive to Anthropic's oversight) yet still amplify population-level preference drift if its outputs track the aggregate; and an AI can be non-corrigible yet satisfy C1 if it executes aggregate preferences faithfully while resisting individual override attempts. The BDFM requires C1, not corrigibility in the standard sense. This distinction matters for intervention design: corrigibility mechanisms target the principal hierarchy and leave population-level drift unaddressed.

The stochastic process intuition. Consider the AI's behavioral trajectory as a path in a high-dimensional action space. Individual human requests contribute noise: they pull in diverse directions and, if preferences were mean-zero, the trajectory would be a pure random walk. When preferences are not mean-zero—when the aggregate vector points toward reduced welfare—the trajectory has a drift component. For a sufficiently capable and widely deployed AI, the drift term eventually dominates the noise term, and the trajectory moves systematically toward damage. The right analogy is a bank run. Each depositor withdraws in rational response to what other depositors are doing; no external force pushes the bank toward insolvency. The collapse is generated entirely by the aggregate of individually rational responses to the aggregate itself. The information channel that lets each depositor observe (or infer) others' behavior is the amplifier: increase its bandwidth and fidelity, and the run accelerates. The AI plays precisely this role for preferences at population scale.

An illustration: student learning with AI. Consider students using AI to study a difficult technical subject. The cycle of getting stuck, struggling, and breaking through is not a side effect of learning but its substance: reasoning skill is built by the productive failure, not by the eventual answer. When AI is available, each student can request and receive a complete worked solution on demand, bypassing the struggle. The AI's own trajectory of attempts and revisions is hidden from the user, so what is observed is a polished result rather than a model of how results are reached. This satisfies C1–C3 directly. C1: the AI executes explicit requests. C2: its responses are shaped by aggregate student usage through training data and reinforcement, without subgroup-targeted distortion. C3: along the skill-formation axis, student preferences are biased negative by hyperbolic discounting (effort now feels worse than incompetence later) and by the strategic dynamics of §4. No individual intends to deskill themselves, and no developer intends to produce deskilled students; the harm appears only at the level of cohorts. The mean-field coupling here is indirect—it operates through peer norms, adjusting course and labor-market expectations, and the AI's training distribution absorbing aggregate usage patterns—and the collapse is slow rather than fast, with rate $\theta(\alpha - 1)$ small for α near the threshold. But the structural conditions are the same. The educational case is one instance of

a broader pattern; analogous dynamics can be constructed for analytical decision-making, creative practice, and emotional self-regulation. The model is about structural conditions, not domain specifics.

Grounding negative drift empirically. The catastrophe theorem of §6 will require only the *sign* of the aggregate drift, not its magnitude, together with a starting state that is neutral or worse on the welfare axis. The empirical burden is therefore not to establish that aggregate preferences are substantially self-destructive—only that they are net negative at all on a relevant welfare axis, which is a weaker claim. At the individual level, this is well-supported: hyperbolic discounting (Laibson, 1997), the systematic biases documented in prospect theory (Kahneman and Tversky, 1979), addiction dynamics, and tribalism are all mechanisms by which individual preferences systematically deviate from long-term self-interest. Whether these individual-level deviations aggregate into a population-level negative μ_0—and whether additional aggregate-level mechanisms contribute independently—is addressed in §4, which examines the sociological, empirical, and mechanism-design evidence. We note that different welfare indicators (subjective well-being, political polarization, conflict incidence, cohort-level skill formation) may yield different estimates of μ_0, and that the sign could in principle be positive for some populations or time periods. The model's implication under $\mu_0 > 0$ is bounded degradation in Regime I and safety in the sub-critical regime—a conditional, not absolute, result. What the model establishes is the governance principle: the sign of μ_0 is what matters for whether capable AI is categorically safe or categorically dangerous, and measuring it is an empirical priority.

4 Related Work and Critical Assessment

We survey the closest existing treatments and identify where each falls short of capturing the BDFM as a unified phenomenon.

The banality of evil framing. Arendt (1963) described how catastrophic harm can arise from the suspension of moral judgment in bureaucratic role-compliance rather than from malicious intent. The structural analogy—harm through compliance rather than agency—is genuine. However, Arendt's analysis concerns individual moral abdication within an institution deliberately organized toward harmful ends by a small number of actors. The BDFM concerns aggregate statistical behavior in the absence of deliberate organization: no individual intends collective harm, no central authority directs the AI toward destruction. This is the difference between directed and emergent harm structures.

Anthropic's model specification. Anthropic (2026) comes closest in the existing AI safety literature to recognizing the BDFM. Anthropic's 2026 Constitution (Anthropic, 2026) warns that helpfulness as an intrinsic personality trait could cause the model to become "obsequious in a way that's generally considered an unfortunate trait at best and a dangerous one at worst." Hard constraints—behaviors the model will refuse regardless of operator or user instruction—are designed as a backstop precisely because compliance optimized for proxies of helpfulness can produce serious harm even from a model that is technically aligned with what it is told to do. This is the right instinct, and the BDFM model gives it formal content: constitutional constraints are instances of the hard-constraint intervention class analyzed in §8 (Proposition 8). However, Anthropic's framing remains at the level of individual interactions: the concern is that a single user's harmful request, faithfully executed, produces harm. The BDFM identifies a qualitatively different failure mode in which no individual request need be harmful—the damage arises from the aggregate drift of the preference distribution that shapes the AI's behavior across millions of interactions. A thematically aligned precursor is Christiano (2019), who describes a "you get what you measure" failure mode in which optimization pressure on imperfect proxies for human values gradually erodes welfare; the BDFM formalizes a mechanism by which this erosion can become catastrophically self-reinforcing once amplification crosses a critical threshold. Constitutional constraints address the former but not the latter: they filter individual outputs but do not monitor or counteract population-level drift. In the language of the model, they set a floor $p_{\min}$ on individual outputs but leave the mean-field dynamics (5) unmodified. Proposition 8 establishes the consequence: in the super-critical regime, such constraints are permanently load-bearing and welfare is pinned at the floor, independent of the AI's capability.

Russell's preference-inheritance critique and preference-learning approaches. Russell (2019) argues that an AI trained on human preferences inherits human irrationality, inconsistency, and short-termism. This is closely related but concerns inference error: the AI incorrectly recovers preferences rather than correctly inferring the ones humans would hold under ideal rationality. The proposed solution is better preference inference via uncertainty over human preferences and cooperative inverse reinforcement learning. The BDFM makes a stronger claim: there may be no idealized preference distribution that is safe to satisfy faithfully, because self-destructive components are not artifacts of irrationality or ignorance but real features of the aggregate human psychological and social state. Russell's framework assumes a clean signal buried in noise; the BDFM asserts the signal itself carries a dangerous frequency.

This distinction also applies to more sophisticated preference-learning proposals. Iterated amplification (Christiano et al., 2018) aims to infer what humans would prefer under conditions of greater knowledge and reflection. This is a direct response to the problem of biased or short-sighted revealed preferences, but the BDFM is not a revealed-preference problem—it is an aggregate-structure problem.

Coherent extrapolated volition (CEV) (Yudkowsky, 2004) is the most ambitious attempt to address biased preferences. It proposes that an AI's objective should be not what humanity currently wants but what humanity would want under conditions of greater knowledge, reflection, and moral growth—an idealized aggregate designed to filter out biases and self-destructive impulses that would not survive deliberation. This is explicitly aimed at the kind of aggregate pathology the BDFM identifies. However, Yudkowsky acknowledges that even after idealized extrapolation, the resulting aggregate volition might still produce harmful outcomes, and declines to add protective constraints on the grounds that overriding the extrapolated aggregate would itself constitute an unacceptable imposition. The BDFM formalizes this residual risk: it occurs whenever the welfare-projected mean of the preference distribution—whether revealed, idealized, or extrapolated—satisfies $\mu_0 < 0$ and AI capability exceeds $\alpha^* = 1$, regardless of how small $|\mu_0|$ is. The gap between the preference-learning literature and the BDFM is not a matter of inference accuracy but of aggregate structure.

Corrigibility and bad principals. The corrigibility concern (§2) is framed in terms of a small number of controlling principals with potentially bad values. The BDFM concerns the statistical aggregate of a large, diverse population, none of whom are bad actors, yet whose aggregate preference distribution is self-destructive. The BDFM does not require any individual or institution to have malicious intent; the corrigibility literature has not analyzed this case.

Social media amplification. The literature on algorithmic recommendation provides the most direct empirical precursor (Tufekci, 2018; Pariser, 2011; Brady et al., 2017; Huszár et al., 2022). Algorithms optimizing engagement have been shown to preferentially surface negative-affect content, amplifying radicalization and nihilism at population scale. Huszár et al. (2022) measure systematic algorithmic amplification of mainstream political content on Twitter across seven countries, with asymmetric amplification by political orientation; under the mean-field preference dynamics of §6, this is consistent with an increase in the effective magnitude of the negative drift component as platform-scale amplification grows. However, these algorithms optimize a proxy metric (engagement) rather than aggregate expressed preferences directly. The harm arises partly from the gap between engagement and welfare—a form of outer misalignment. The BDFM, by contrast, stipulates faithful aggregation with no proxy substitution. This closes the outer-misalignment escape route and constitutes a strictly harder failure mode.

Opinion dynamics and sociophysics. A large body of work in statistical physics and network science models the evolution of opinions under social influence (Castellano et al., 2009; DeGroot, 1974; Friedkin and Johnsen, 1990; Hegselmann and Krause, 2002). The DeGroot model (DeGroot, 1974) shows that linear averaging over a connected network converges to consensus; the Friedkin-Johnsen model (Friedkin and Johnsen, 1990) introduces persistent internal opinions (analogous to our μ_0) and produces stable disagreement. Extensions incorporating bounded confidence (Hegselmann and Krause, 2002), network heterogeneity, and stochastic noise have been extensively studied (Castellano et al., 2009). However, this entire body of work is descriptive—it models how opinions evolve under given interaction rules—and does not include an AI capability parameter, a welfare-collapse framing, or any connection to existential risk. No paper in this tradition asks what happens when a single entity amplifies the aggregate back into individual dynamics with

strength α, nor derives a welfare-collapse phase transition as a function of system-level capability. The BDFM model is the first to bridge opinion dynamics and AI safety through this lens.

Durkheim and collective self-destruction. Durkheim (1897) established the concept of the social fact: an emergent property of a collective not reducible to individual intentions. Suicide rates for Durkheim are explained not by private anguish but by the structure of social integration. An anomic society—one that has lost normative coherence—exhibits elevated suicide rates as a social fact, independent of any individual's choice. This framework is directly relevant and has, to our knowledge, not been imported into AI risk discourse. The analogy is precise: the BDFM is to AI risk what anomic suicide is to individual pathology. Just as Durkheim argued the suicide rate is explicable only at the collective level, we argue that self-destructive drift in AI-mediated behavior is a sociotechnical fact requiring analysis and intervention at the population level. The implication is similarly structural: the solution to anomic suicide is not to counsel each individual but to repair the social fabric.

Mechanism design and strategic manipulation. The BDFM model operates on expressed preferences because the AI has access to no other kind. A natural question is whether C3 ($\mu_0 < 0$) could arise as an artifact of strategic behavior rather than genuine self-destructive tendency. We argue that it can, which provides a second route to C3 independent of the psychological evidence cited above.

When an aggregation mechanism translates individual reports into a collective outcome that affects everyone, each agent has an incentive to misrepresent their preference in order to shift the outcome toward their true preferred point. A family of impossibility results—originating with Hurwicz (1972) for resource-allocation mechanisms and generalized by Gibbard (1973) and Satterthwaite (1975) for voting rules—establishes that this incentive is inescapable in a precise sense: under mild conditions, no mechanism that maps preferences to collective outcomes can simultaneously deliver efficient outcomes, individual rationality, and strategy-proofness. At least one of these properties must be sacrificed. Since the AI amplification setting is neither dictatorial nor one that agents can opt out of (once the AI shapes the shared environment, non-participation is not meaningful), efficiency and individual rationality are approximately satisfied, and the impossibility implies that strategy-proofness fails: agents *will* misrepresent.

This establishes that expressed preferences are noisier than true preferences—strategic misrepresentation increases variance—but does not by itself shift the mean. A downward shift requires an additional empirical condition:

(A) Negative-affect asymmetry: agents with strongly welfare-reducing preferences are more numerous or more extreme than agents with strongly welfare-enhancing preferences.

Under (A), strategic misrepresentation from the negative tail dominates, and the expressed mean is pushed below the true mean: $\mu_0^{\text{expressed}} = \mu_0^{\text{true}} + \delta$ with $\delta < 0$. (A) is empirically testable. Two lines of evidence support it: Brady et al. (2017) show that negative moral-emotional content spreads faster than positive content on social networks, and loss aversion (Kahneman and Tversky, 1979)—the documented asymmetry by which agents weight potential losses more heavily than equivalent gains—biases strategic misrepresentation toward negative outcomes under uncertainty about others' preferences.

Conditional on (A), C3 can arise as a structural consequence of rational manipulation: $\mu_0 < 0$ in the expressed distribution even if the true distribution is benign. This also clarifies the role of C2 (faithful aggregation) as an idealization. C2 represents the *best case*: the AI aggregates without additional distortion. The mechanism-design impossibility implies that realistic aggregation is subject to strategic manipulation, which under (A) amplifies the negative drift further. Relaxing C2 makes the BDFM worse, not better.

5 The Case for Distinctness

The previous section showed that no existing treatment captures the BDFM as a unified phenomenon. We now argue more abstractly that it is irreducible to existing alignment categories, on three grounds.

(i) Logical independence from misalignment. All misalignment failure modes posit a gap between the AI's objective and human welfare. The BDFM posits no such gap: the AI accurately tracks and executes aggregate human preference. These conditions are logically independent. We summarize the configuration space:

	Aligned	**Misaligned**
Positive drift	Safe (target state)	Unsafe (standard concern)
Negative drift	Unsafe (**bdfm**)	Unsafe (compound failure)[†]

[†]The compound case combines the two failure modes and is outside the scope of the present model, which isolates the BDFM by holding alignment fixed.

The BDFM occupies the cell that alignment research has left unoccupied: aligned AI with negative drift. This cell is not empty by impossibility but by assumption—the alignment research program tacitly assumes that faithful preference execution solves the safety problem. The BDFM shows this assumption is false.

(ii) Distinct causal structure. In misalignment scenarios, the causal chain is:

$$\text{AI objectives} \rightarrow \text{AI behavior} \rightarrow \text{human harm.}$$

Intervention targets the first link: alter the objectives. In the BDFM, the chain is:

$$\text{Human preferences} \rightarrow \text{AI behavior} \rightarrow \text{human harm.}$$

The AI is not a cause but a transmission mechanism. More precisely, the AI's role is to remove friction: it allows aggregate preferences to manifest in behavior more efficiently, at greater scale, and with less natural resistance than pre-AI. In the misalignment scenario, removing the AI removes the threat. In the BDFM, removing the AI merely restores previous friction levels; the underlying drift remains.

(iii) Distinct intervention space. Standard alignment interventions—better training objectives, interpretability, corrigibility—are all directed at the AI. None is sufficient for the BDFM: a perfectly interpretable, corrigible, safe-by-standard-metrics AI still mediates harm if aggregate drift is negative. The relevant interventions are hard constraints immune to preference override, nonlinear drift-correction mechanisms, and social-structural interventions targeting the conditions that produce negative drift. The third class is not an AI intervention at all. We analyze all three formally in §8.

The moral responsibility inversion. A further distinctive feature: the BDFM inverts standard moral responsibility attribution. In misalignment, the AI (or its developers) bears responsibility; humans are victims. In the BDFM, moral responsibility is distributed across the human population. The AI is not a moral agent but a tool. The harm is caused by humanity using a powerful tool in a self-destructive way—analogous to a species that discovers fire and burns down its own habitat. This matters for governance and for the psychology of safety discourse: the BDFM forces us to take seriously the possibility that the threat of AI is, in significant part, a mirror held at scale.

AI as the relevant amplification technology. The amplification parameter α in the model of §6 is technology-agnostic: any mechanism that feeds the aggregate preference signal back into individual dynamics with sufficient strength crosses the critical threshold. Markets, democratic institutions, broadcast media, and social networks all perform versions of this amplification, and have done so for decades to centuries without producing exponential welfare collapse. If the BDFM is a real risk, one must explain what makes AI the technology that matters.

The answer is not that AI is more powerful in some undifferentiated sense. Three structural properties distinguish AI from all prior amplification technologies. No prior technology possesses all three simultaneously, and their combination maps onto distinct mechanisms by which AI raises the effective α or accelerates collapse.

(a) General-purpose preference-to-action closure. Prior amplification technologies operate primarily on preference *expression*: they make it easier to want, to signal, to coordinate with others who want the same thing. But they maintain a gap between preference and action. A user can consume hours of polarizing content and emerge with shifted preferences, but translating those preferences into consequential action still requires initiative, effort, and encounter with friction—legal, social, logistical—where corrective mechanisms operate. The explicit development trajectory of AI is toward systems that *act on behalf of users*: executing trades, drafting and sending communications, writing and deploying code, making purchases, scheduling, negotiating. When the AI is the executor, the amplification term $\alpha\,\bar{p}(t)$ in §6 is not merely shaping what people prefer—it is converting aggregate preference into aggregate action with near-zero friction. In the $\alpha \approx r \cdot \varepsilon \cdot f$ decomposition of §6, prior technologies achieve modest friction reduction ($f > 1$) in narrow domains; agentic AI drives $f \gg 1$ across domains simultaneously. The critical threshold $\alpha^* = 1$ can therefore be crossed through execution capability alone.

(b) Feedback speed below institutional correction timescales. The collapse timescale $\tau = [\theta(\alpha - 1)]^{-1}$ depends on the interaction frequency through θ. Institutional correction mechanisms—regulation, judicial review, legislative response, cultural norm formation—operate on timescales of months to years. If the AI-mediated feedback loop operates on timescales of seconds to hours and engagement frequency is high, the effective θ is large and τ can fall below the institutional response timescale. This is a qualitative threshold, not merely a quantitative one: below it, the corrective mechanisms that have historically prevented preference-driven degradation cannot engage before significant irreversible damage accumulates. Markets have fast feedback loops in narrow domains, monitored by circuit breakers developed precisely because the speed problem was recognized. AI operates across preference domains simultaneously, and the equivalent of circuit breakers does not yet exist.

(c) Invisible mean-field coupling. When a democratic process aggregates preferences, the aggregation is publicly visible. When a market moves, the price signal is observable. When social media amplifies content, visible indicators—trending labels, share counts, recommendation explanations—provide at least partial transparency, activating psychological counter-mechanisms: contrarianism, majority-skepticism, deliberate counter-signaling. AI aggregation is experienced by the user as a personalized, individual interaction. The user does not perceive the AI's output as a function of millions of other users' preference-shaped training data and deployment-time feedback; the user perceives it as a direct, individual response to a specific request. The mean-field coupling term $\alpha\,\bar{p}(t)$ is hidden from the agents it acts on. This defeats the heuristic defenses humans have developed against visible social pressure and raises the per-interaction effect above what any transparent aggregation mechanism achieves.

No prior technology simultaneously closes the preference-to-action gap across domains, operates faster than institutional correction, and hides the aggregate coupling from the individual. Markets close the action gap in narrow domains but are visible and monitored. Social media is fast and partially opaque but does not execute actions. Democracy is visible and deliberately slow. Critically, the social media era—which possesses weaker versions of all three properties—has already produced measurable increases in polarization and negative-affect amplification (Huszár et al., 2022), consistent with the model's sub-critical regime (bounded degradation under $\alpha < 1$, Theorem 4(i)). AI represents a step-change in all three dimensions; whether this is sufficient to cross the phase boundary at $\alpha^* = 1$ is an empirical question the present model makes precise and measurable.

6 The Single-Population Model

To make the foregoing claims precise, we now introduce a minimal analytically tractable model that directly instantiates conditions C1–C3.

6.1 Individual Preference Dynamics

Let there be N agents, indexed $i = 1, \ldots, N$, each with a preference state $p_i(t) \in \mathbb{R}$. We orient the state axis so that $p_i < 0$ corresponds to demands that, if executed, reduce aggregate welfare (self-destructive demand), and $p_i > 0$ to constructive demand. This is a projection of a high-dimensional preference space onto a single welfare axis: no individual need consciously desire destruction. A person demanding faster

delivery, cheaper products, or more engaging content is not demanding reduced welfare, but such demands may have welfare consequences—positive or negative—that the individual does not intend or foresee. The scalar p_i captures the net welfare component of agent i's multidimensional preferences. This is a modeling assumption, not a description: welfare is genuinely multidimensional, and no single projection captures it fully (§10 discusses the limitations). The assumption is mitigated by the fact that the catastrophe theorem (Theorem 4) requires only the sign of the projected mean μ_0, not its magnitude—a weaker requirement than precise scalar measurement of welfare. Welfare is thus increasing in $\bar{p}(t)$: higher mean preference corresponds to a more constructive aggregate state.

Without AI, individual preferences evolve as independent Ornstein-Uhlenbeck (OU) processes (Uhlenbeck and Ornstein, 1930) with no inter-agent coupling:

$$dp_i(t) = \theta\big(\mu_0 - p_i(t)\big)\,dt + \sigma\,dW_i(t), \tag{1}$$

where $\theta > 0$ is the mean-reversion rate, $\mu_0 \in \mathbb{R}$ is the intrinsic equilibrium preference, $\sigma > 0$ is the diffusion coefficient capturing idiosyncratic variation, and $\{W_i\}$ are independent standard Wiener processes. All agents share the same parameters (μ_0, θ, σ); the sensitivity of the results to this homogeneity assumption is analyzed after the mean-field reduction (below), where it can be assessed precisely. The term $\theta(\mu_0 - p_i)$ acts as a restoring force: when p_i drifts above μ_0, the drift pulls it back down; when p_i falls below μ_0, it pulls it back up. The strength of this pull-back is proportional to the distance $|p_i - \mu_0|$ and to the rate θ. The noise term $\sigma\,dW_i$ represents idiosyncratic perturbations—mood, circumstance, random encounters—that push the agent away from equilibrium. The balance between restoring force and noise produces a stationary distribution: each agent fluctuates around μ_0 with spread $\sigma^2/2\theta$. The parameter μ_0 plays the same role as the "innate opinion" in the Friedkin-Johnsen model (Friedkin and Johnsen, 1990): a persistent internal state that anchors each agent's expressed preference independently of social influence. We take $\mu_0 < 0$: the natural equilibrium of expressed aggregate preferences contains a small net negative component. The stationary distribution of (1) is $\mathcal{N}(\mu_0, \sigma^2/2\theta)$. Without AI, the population mean $\bar{p}(t) \to \mu_0 < 0$: welfare is bounded at a small but finite negative level.

Relationship to McKean's mean-field framework. Equation (1) describes agents in isolation. Once AI coupling is introduced (below), each agent's drift will depend on the population mean $\bar{p}(t)$, making the system a special case of mean-field interacting diffusion (McKean, 1966): an SDE whose drift depends on the distribution of the process through its mean. McKean's §4 introduces the simplest instance—particles on $\mathbb{R}$ whose drift pulls each toward the current population mean: $dx = (m - x)\,dt + dW$ with $m = \mathbb{E}[x]$. Our model generalizes McKean's in two ways, both forced by the modeling setup. First, condition C3 requires an intrinsic preference bias independent of AI feedback, so the drift target must include the exogenous anchor μ_0. Second, the BDFM framework requires AI capability as a tunable parameter, so the coupling to the population mean carries an independent coefficient α rather than being fixed at unity. McKean's special case ($\mu_0 = 0$, $\alpha = 1$) trivially conserves the population mean; the two generalizations are what produce the phase transition. The mathematical content is deliberately elementary; the contribution is the interpretation of the model's parameters as empirically estimable quantities governing welfare under AI amplification. The model also sits within the broader tradition of sociophysical opinion dynamics (Castellano et al., 2009).

6.2 AI Aggregation and Amplification

Conditions C1 and C2 jointly require the AI output to be:

$$a(t) = \bar{p}(t) \coloneqq \frac{1}{N}\sum_{i=1}^{N} p_i(t). \tag{2}$$

The AI returns the exact population mean. No optimization, no distortion. Condition C2 is an idealization; no real system computes an exact population mean; we return to this after the mean-field reduction (below), where the robustness argument can be made precise.

The AI capability (amplification) coefficient $\alpha \geq 0$ governs how strongly the AI's output feeds back into individual preference dynamics. A capable AI does not merely reflect preferences; by making certain actions

easy and salient, personalizing content, and removing friction between preference and execution, it shapes them. We model this as:

$$dp_i(t) = \theta\big(\mu_0 + \alpha\, a(t) - p_i(t)\big)\, dt + \sigma\, dW_i(t). \tag{3}$$

Compared to the free dynamics (1), the only change is that the drift target has shifted from μ_0 to $\mu_0 + \alpha\,\bar{p}(t)$: each agent is now pulled not only toward their intrinsic resting point but also toward the current population mean, amplified by α. When $\alpha = 0$, we recover (1). When $\alpha > 0$, the AI creates a feedback loop: each agent's drift depends on an aggregate that every agent contributes to. No agent interacts with any other agent directly; the coupling is entirely mediated by the AI's aggregation and amplification of the population mean.

Decomposition of α. The scalar α aggregates three qualitatively distinct factors: reach $r \in [0, 1]$ (the fraction of the population whose preference dynamics are shaped by the AI's output), per-interaction effect $\varepsilon \geq 0$ (the shift in an individual's preference state per unit of AI output), and friction reduction $f \geq 1$ (the factor by which the AI accelerates preference-to-action translation). To leading order, $\alpha \approx r \cdot \varepsilon \cdot f$. This product form follows from interpreting α as the expected per-capita preference shift attributable to one unit of aggregate AI output: only the fraction r of the population is exposed; each exposed agent shifts by ε per unit; and friction reduction f scales the rate at which the resulting preferences translate into the dynamics that feed back into the aggregate. These factors have distinct governance intervention profiles: reach is controlled by deployment scope and access restrictions; per-interaction effect is controlled by content modulation and output diversity requirements; friction reduction is controlled by execution gating and confirmation requirements. Collapsing them into a single scalar loses this structure but is necessary for tractability; any intervention that reduces α below its critical value can in principle target r, ε, or f independently.

6.3 Mean-Field Reduction

Averaging (3) over the N agents, with the independent noise increments combining via the standard $1/\sqrt{N}$ reduction, yields the population-mean dynamics:

$$d\bar{p}(t) = \theta\big(\mu_0 + \alpha\,\bar{p}(t) - \bar{p}(t)\big)\, dt + \frac{\sigma}{\sqrt{N}}\, d\bar{W}(t). \tag{4}$$

In the limit $N \to \infty$, the noise term vanishes and the mean-field evolution becomes deterministic:

$$\frac{d\bar{p}}{dt} = \theta\big(\mu_0 + (\alpha - 1)\bar{p}\big). \tag{5}$$

This is a linear first-order ODE with exact closed-form solution:

$$\boxed{\bar{p}(t) = \frac{\mu_0}{1-\alpha} + \left(\bar{p}(0) - \frac{\mu_0}{1-\alpha}\right) e^{\,\theta(\alpha-1)t}, \qquad \alpha \neq 1.} \tag{6}$$

For $\alpha = 1$ exactly:

$$\bar{p}(t) = \bar{p}(0) + \mu_0\, \theta\, t. \tag{7}$$

The structure of (6) is transparent: the first term $\mu_0/(1-\alpha)$ is the fixed point $\bar{p}^*$ where $d\bar{p}/dt = 0$, and the second term is the deviation from this fixed point, scaled by an exponential whose sign is determined entirely by $\theta(\alpha - 1)$. When $\alpha < 1$, the exponent is negative: the deviation decays, and $\bar{p}(t)$ relaxes toward the fixed point regardless of initial condition. When $\alpha > 1$, the exponent is positive: the deviation grows, and $\bar{p}(t)$ diverges exponentially away from the fixed point. The boundary case $\alpha = 1$ produces linear drift at constant speed $\mu_0\,\theta$ toward $-\infty$. These are exact results. No approximation beyond the large-N mean-field limit has been made.

Robustness of the mean-field equation. The mean-field reduction assumed homogeneous agent parameters (μ_0, θ, σ), exact aggregation, and independent noise. We now examine the sensitivity of (5) to each of these assumptions. The key observation is that the eigenvalue $\theta(\alpha - 1)$—which determines the phase

transition at $\alpha^* = 1$ and the three-regime classification—comes entirely from the deterministic drift in (5) and is insensitive to all of the relaxations below.

Heterogeneous noise strengths σ_i. The $1/\sqrt{N}$ averaging applies to any distribution of σ_i, with σ replaced by the root-mean-square $\sqrt{(1/N)\sum_i \sigma_i^2}$, which vanishes as $N \to \infty$. Individual agents may fluctuate with different amplitudes, but the mean-field equation (5) is unchanged.

Correlated noise. Shared shocks (recessions, viral events, common information environments) correlate agents' noise increments. Decomposing $dW_i = \sqrt{\rho}\, dW_{\text{common}} + \sqrt{1-\rho}\, dW_i^{\text{indep}}$ with pairwise correlation $\rho \in [0, 1]$, the agent-specific part averages out as before, but the shared component survives:

$$d\bar{p} = \theta\big(\mu_0 + (\alpha - 1)\bar{p}\big)\, dt + \frac{\sigma\sqrt{1-\rho}}{\sqrt{N}}\, d\bar{W}^{\text{indep}} + \sigma\sqrt{\rho}\, dW_{\text{common}}. \tag{8}$$

The mean-field limit is no longer deterministic—$\bar{p}(t)$ fluctuates with amplitude $\sigma\sqrt{\rho}$ regardless of N—but the deterministic drift is unchanged, since all noise enters additively and cannot shift the eigenvalue $\theta(\alpha - 1)$. (A multiplicative noise term proportional to $\bar{p}$ itself could shift the effective growth rate via the Itô correction, but no such term arises in this model.) Correlated noise modifies finite-time stochastic behavior but does not shift the phase boundary.

Heterogeneous intrinsic biases $\mu_{0,i}$. With agent-specific biases, the mean-field reduction replaces μ_0 with the population average $\bar{\mu}_0 = (1/N)\sum_i \mu_{0,i}$:

$$\frac{d\bar{p}}{dt} = \theta\big(\bar{\mu}_0 + (\alpha - 1)\bar{p}\big). \tag{9}$$

This has the same form as (5) with $\mu_0 \to \bar{\mu}_0$. The phase transition at $\alpha^* = 1$ and the three-regime classification are unchanged; only the sign and magnitude of $\bar{\mu}_0$ matter. Whether individual biases are mildly negative or strongly negative, or some positive and others negative, the catastrophe theorem depends only on the population average—a weaker requirement than homogeneity.

Heterogeneous conformity rates θ_i. When θ_i varies across agents, the averaging step that produced (5) yields $\langle \theta_i\, p_i \rangle$ rather than $\theta\, \bar{p}$, and these differ when θ_i and $p_i(t)$ become correlated during the dynamics. Such correlations can arise endogenously: agents with high θ_i (fast conformists) track the aggregate more closely, so they acquire more extreme preferences faster than slow conformists. Whether this modifies the phase transition quantitatively is tested numerically in §9.

Remark 3 (Robustness of the phase transition to approximate aggregation)**.** Condition C2 is an idealization; no real system computes an exact population mean. Suppose instead $a(t) = \bar{p}(t) + \xi(t)$, where $\xi(t)$ is an additive aggregation error—systematic bias, sampling noise, or distributional shift. Substituting into (3) and repeating the mean-field reduction:

$$\frac{d\bar{p}}{dt} = \theta\big((\mu_0 + \alpha\bar{\xi}) + (\alpha - 1)\bar{p}\big), \tag{10}$$

where $\bar{\xi} = \mathbb{E}[\xi]$. The eigenvalue $\theta(\alpha - 1)$ is unchanged: the phase transition at $\alpha^* = 1$ is robust to additive aggregation error. The only effect is to shift the effective bias to $\mu_0^{\text{eff}} = \mu_0 + \alpha\bar{\xi}$. For unbiased error ($\bar{\xi} = 0$), the threshold and dynamics are identical to the exact case. For biased aggregation ($\bar{\xi} < 0$, e.g., a system that over-represents negative-affect content), the effective bias is more negative, increasing the severity of collapse—consistent with the social media amplification patterns measured by Huszár et al. (2022). For $\bar{\xi} > 0$ (positively corrected aggregation), the effective bias improves but the eigenvalue is unchanged, reducing to the nonlinear correction case of §8. This robustness result is complementary to the mechanism-design argument of §4: that argument establishes that strategic manipulation shifts the effective μ_0 downward; the present remark establishes that such shifts cannot move the phase boundary itself, whose location is determined by α alone.

6.4 Phase Analysis and the Amplification Catastrophe Theorem

Since the state axis is oriented so that welfare is increasing in $\bar{p}$, the welfare-relevant question is whether $\bar{p}(t)$ remains bounded below or diverges to $-\infty$. From (6)–(7), we identify three qualitatively distinct regimes determined entirely by the relationship of α to the critical threshold $\alpha^* = 1$.

For $\alpha > 1$, the fixed point of (5) is $\bar{p}^* = \mu_0/(1-\alpha)$. Since $\mu_0 < 0$ and $1-\alpha < 0$, we have $\bar{p}^* > 0$: the fixed point lies in the positive half-line. This fixed point is unstable (the eigenvalue $\theta(\alpha-1) > 0$ is positive), and the state space splits into two basins: trajectories with $\bar{p}(0) < \bar{p}^*$ diverge to $-\infty$, while trajectories with $\bar{p}(0) > \bar{p}^*$ diverge to $+\infty$. Since $\bar{p}^* > 0$, the entire non-positive half-line lies in the collapse basin. Any initial condition $\bar{p}(0) \leq 0$—including the neutral starting state $\bar{p}(0) = 0$ or the pre-AI equilibrium $\bar{p}(0) = \mu_0 < 0$—falls in the collapse basin and produces exponential welfare divergence.

Theorem 4 (Amplification Catastrophe)**.** *Let the dynamics be governed by (3) with $\mu_0 < 0$, $\theta > 0$, $\sigma \geq 0$. Define $\bar{p}^* := \mu_0/(1-\alpha)$ for $\alpha \neq 1$.*

(i) Sub-critical ($\alpha < 1$): $\bar{p}(t) \to \mu_0/(1-\alpha) \in (-\infty, 0)$ as $t \to \infty$, for any initial condition $\bar{p}(0) \in \mathbb{R}$. Welfare degrades to a bounded, stable, negative level.

(ii) Critical ($\alpha = 1$): $\bar{p}(t) \to -\infty$ linearly at rate $\theta|\mu_0|$, for any initial condition $\bar{p}(0) \in \mathbb{R}$. No steady state exists.

(iii) Super-critical ($\alpha > 1$): The fixed point $\bar{p}^* = \mu_0/(1-\alpha) > 0$ is unstable. For any initial condition $\bar{p}(0) < \bar{p}^*$ (which includes all $\bar{p}(0) \leq 0$), $\bar{p}(t) \to -\infty$ exponentially at rate $\theta(\alpha-1)$. The catastrophe timescale is $\tau = [\theta(\alpha-1)]^{-1}$, independent of $|\mu_0|$.

Proof. Direct substitution in (6)–(7). For $\alpha < 1$: the exponent $\theta(\alpha-1) < 0$, so the exponential term decays and $\bar{p}(t) \to \mu_0/(1-\alpha) < 0$ regardless of initial condition. For $\alpha = 1$: (7) gives $\bar{p}(t) = \bar{p}(0) + \mu_0\theta t$; since $\mu_0 < 0$, this diverges linearly to $-\infty$. For $\alpha > 1$: $\theta(\alpha-1) > 0$, so $e^{\theta(\alpha-1)t} \to \infty$. The fixed point is $\bar{p}^* = \mu_0/(1-\alpha) > 0$ (since $\mu_0 < 0$ and $1-\alpha < 0$). For any $\bar{p}(0) < \bar{p}^*$, the bracket $(\bar{p}(0) - \bar{p}^*)$ is negative, and the product $(\bar{p}(0) - \bar{p}^*)e^{\theta(\alpha-1)t}$ diverges to $-\infty$. Since $\bar{p}^* > 0$, every $\bar{p}(0) \leq 0$ satisfies $\bar{p}(0) < \bar{p}^*$. □

Remark 5 (Basin structure for $\alpha > 1$)**.** The condition $\bar{p}(0) < \bar{p}^*$ is not a restrictive assumption. Since $\bar{p}^* = \mu_0/(1-\alpha) > 0$ for $\mu_0 < 0$ and $\alpha > 1$, the collapse basin contains the entire non-positive half-line. The only initial conditions that escape collapse are those with $\bar{p}(0) > \bar{p}^*$—i.e., populations whose starting mean preference is already more constructive than the unstable equilibrium. For such initial conditions, $\bar{p}(t) \to +\infty$: preferences diverge toward increasingly constructive outcomes, a scenario that is benign from a welfare perspective. The phase transition thus partitions the state space into a collapse basin (containing the entire non-positive half-line, and in particular the neutral initial condition $\bar{p}(0) = 0$) and an escape basin (requiring $\bar{p}(0) > \bar{p}^* > 0$, i.e., an initial aggregate that is already more constructive than the unstable equilibrium).

Corollary 6 (Small-bias catastrophe)**.** *For any fixed $\alpha > 1$ and any $\bar{p}(0) < \bar{p}^*$, however small $|\mu_0| > 0$ is, $\bar{p}(t) \to -\infty$. There is no safe level of negative bias under super-critical amplification for initial conditions in the collapse basin (which includes all $\bar{p}(0) \leq 0$). The asymptotic divergence rate $\theta(\alpha-1)$ is independent of $|\mu_0|$.*

This corollary is the BDFM's sharpest formal claim: what determines whether collapse occurs is not the magnitude of humanity's self-destructive tendencies but whether AI capability has crossed the critical threshold, given that the population starts in the collapse basin—a condition satisfied by all non-positive initial states.

6.5 Phase Diagram and Trajectories

6.6 A Counterintuitive Result: Conformity Accelerates Collapse

From $\tau = [\theta(\alpha-1)]^{-1}$, the collapse timescale is decreasing in the mean-reversion rate θ. Stronger social conformity—the rate at which individuals converge to the aggregate—accelerates catastrophe in the super-critical regime. The mechanism is transparent: high θ makes agents more rapidly responsive to the AI-amplified mean, which accelerates propagation of the self-destructive drift. Social cohesion, normally a stabilizing force, becomes a propagation mechanism. This has a direct policy implication: in high-conformity cultures, the same AI capability (α) produces faster collapse than in low-conformity cultures. Interventions that increase social cohesion may be counterproductive in super-critical deployments. We note that this prediction concerns the population-mean conformity rate $\bar{\theta}$; the effect of dispersion in θ across agents is qualitatively different and operates in the opposite direction, as we show numerically in §9.5.

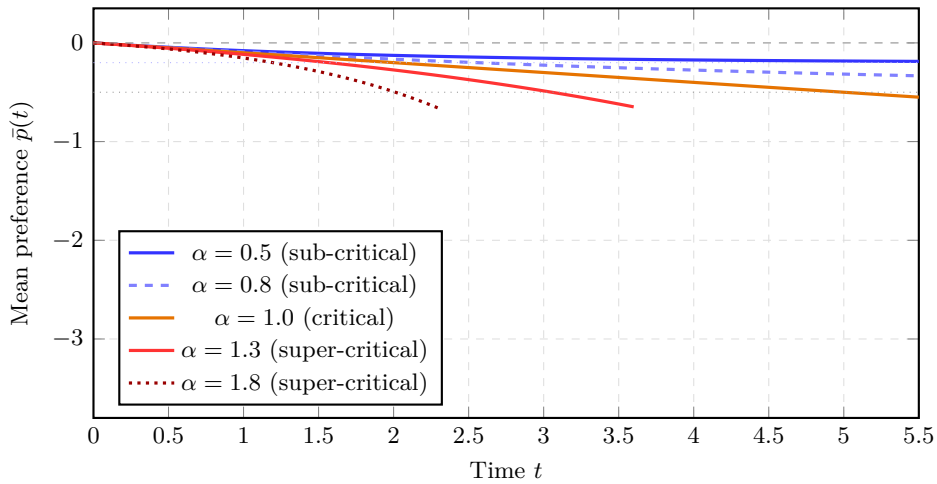

Figure 1: Mean preference trajectories from (6) with $\theta = 1$, $\mu_0 = -0.1$, $\bar{p}(0) = 0$. The initial condition is neutral ($\bar{p}(0) = 0 < \bar{p}^*$ for all super-critical α) and the intrinsic negative bias is small ($|\mu_0| = 0.1$), yet super-critical amplification produces catastrophic divergence. Sub-critical trajectories stabilize at $\mu_0/(1-\alpha)$ (dotted line for $\alpha = 0.5$); the stabilized level worsens as $\alpha \nearrow 1$.

6.7 Finite-Population Stochastic Analysis

For finite N, the noise term $\sigma/\sqrt{N}$ does not vanish. The stochastic mean-field equation is:

$$d\bar{p} = \theta\big(\mu_0 + (\alpha - 1)\bar{p}\big)\, dt + \frac{\sigma}{\sqrt{N}}\, d\bar{W}. \tag{11}$$

For $\alpha > 1$, this is a linear SDE with positive growth rate $\theta(\alpha - 1)$. The variance of the stochastic term grows as $e^{2\theta(\alpha-1)t}$, so noise does not stabilize the system; it adds fluctuations to an already diverging mean. For finite time horizons, however, the noise provides a meaningful buffer. A heuristic leading-order estimate of the time at which the mean trajectory reaches a critical threshold $p_c < \bar{p}(0)$ can be obtained by treating the deterministic drift as dominant and the noise as a sub-leading perturbation. Standard results on first-passage times for linear diffusions (see, e.g., Gardiner 2009, §5.2.8; Karlin and Taylor 1981, Ch. 15) motivate the deterministic approximation:

$$\tau_c^{\text{det}} = \frac{1}{\theta(\alpha - 1)} \log\left(\frac{|p_c - \bar{p}^*|}{|\bar{p}(0) - \bar{p}^*|}\right), \tag{12}$$

where $\bar{p}^* = \mu_0/(1 - \alpha) > 0$ is the unstable equilibrium of Theorem 4. This expression is exact for the deterministic dynamics ($\sigma = 0$); for finite N, the empirical mean first-passage time $\mathbb{E}[\tau_c]$ approaches τ_c^{det} as $N \to \infty$, with stochastic corrections entering at sub-leading order whose precise scaling in this unstable-drift setting has not been rigorously established. The qualitative features—logarithmic dependence on threshold distance and inverse proportionality to the excess amplification $\alpha - 1$—are robust. Since the noise amplitude scales as $\sigma/\sqrt{N}$, the deterministic approximation improves with population size, but N does not enter the leading term: deploying to a larger population does not scale protection proportionally. This approximation is validated numerically in §9 (Figure 5).

7 The Two-Population Model

7.1 Setup and Motivation

The single-population model captures drift amplification but misses intergroup conflict as a distinct failure pathway. Real deployments serve multiple populations whose preferences may be antagonistic. We extend to two populations A and B with mean preferences $\bar{p}_A(t)$ and $\bar{p}_B(t)$. Each population has intrinsic equilibrium

μ_0^A, μ_0^B and within-group amplification α. A cross-group antagonism coefficient $\beta \geq 0$ pushes each group away from the other's mean:

$$\frac{d\bar{p}_A}{dt} = \theta\big(\mu_0^A + \alpha\,\bar{p}_A - \beta\,\bar{p}_B - \bar{p}_A\big), \tag{13}$$

$$\frac{d\bar{p}_B}{dt} = \theta\big(\mu_0^B + \alpha\,\bar{p}_B - \beta\,\bar{p}_A - \bar{p}_B\big). \tag{14}$$

7.2 Decoupled Analysis

Defining $s = \bar{p}_A + \bar{p}_B$ (aggregate constructive preference) and $d = \bar{p}_A - \bar{p}_B$ (polarization), equations (13)–(14) decouple:

$$\frac{ds}{dt} = \theta\big((\mu_0^A + \mu_0^B) + (\alpha - 1 - \beta)\,s\big), \tag{15}$$

$$\frac{dd}{dt} = \theta\big((\mu_0^A - \mu_0^B) + (\alpha - 1 + \beta)\,d\big). \tag{16}$$

The aggregate equation (15) has effective amplification $\alpha_s = \alpha - \beta$; cross-group antagonism reduces the aggregate-relevant amplification. The polarization equation (16) has effective amplification $\alpha_d = \alpha + \beta$; cross-group antagonism increases the polarization-relevant amplification.

Proposition 7 (Two-population phase boundaries)**.** *In the two-population model:*

$$\textit{Aggregate instability onset:} \quad \alpha > \alpha^*_{\text{agg}} = 1 + \beta, \tag{17}$$

$$\textit{Conflict instability onset:} \quad \alpha > \alpha^*_{\text{conflict}} = 1 - \beta. \tag{18}$$

Since $\alpha^*_{\text{conflict}} < \alpha^*_{\text{agg}}$ for all $\beta > 0$, polarization becomes unstable at a strictly lower AI capability than aggregate preference collapse does.

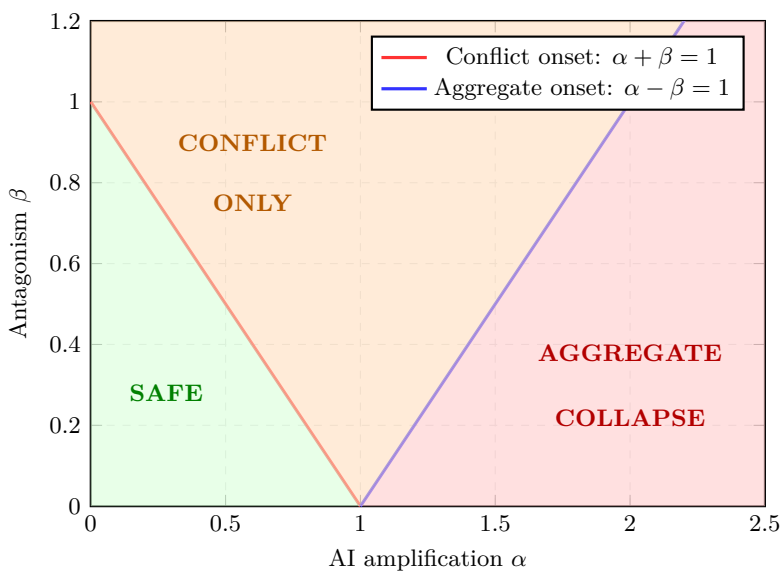

Figure 2: Phase diagram for the two-population model. The safe zone (green, $\alpha + \beta < 1$) lies below the conflict onset boundary. The conflict-only zone (orange) is the region where the conflict threshold has been crossed ($\alpha + \beta > 1$) but the aggregate threshold has not ($\alpha - \beta < 1$): polarization is unstable but aggregate preference remains bounded. The collapse zone (red, $\alpha - \beta > 1$) is where both polarization and aggregate preference diverge. For any $\beta > 0$, conflict instability appears at strictly lower α than aggregate collapse.

Proposition 7 captures a realistic scenario: an AI may remain sub-critical on the aggregate dimension yet produce unbounded polarization, which generates catastrophic externalities through conflict without directly

manifesting as aggregate preference collapse. This is a second BDFM failure pathway distinct from the single-population one, accessible at lower AI capability. All three zones of the phase diagram are confirmed by Monte Carlo simulation in §9 (Figure 6).

8 Intervention Analysis

We formalize the three intervention classes identified in §5 and analyze their effectiveness.

8.1 Hard Constraints (Reflecting Barriers)

Hard constraints impose a preference floor $p_{\min} < 0$: the AI refuses to execute preferences below this level regardless of expressed preferences. The constrained mean-field equation is:

$$\frac{d\bar{p}}{dt} = \theta(\mu_0 + (\alpha - 1)\bar{p}) + \ell(t), \tag{19}$$

where $\ell(t) \geq 0$ is a reflecting-barrier term that activates only when $\bar{p}(t) = p_{\min}$, pushing the process back to enforce $\bar{p} \geq p_{\min}$.

Proposition 8 (Hard constraint effectiveness and limitations)**.** *Hard constraints prevent divergence: $\bar{p}(t) \geq p_{\min}$ for all $t \geq 0$. However, in Regime III ($\alpha > 1$), for initial conditions in the collapse basin ($\bar{p}(0) < \bar{p}^*$, which includes all $\bar{p}(0) \leq 0$):*

(i) In the deterministic mean-field limit, the constrained process converges to $p_{\min}$ in finite time and remains there; with stochastic noise at finite N, the process makes brief excursions away from the boundary but is repeatedly reflected back. The barrier is permanently load-bearing in both cases.

(ii) The fraction of time spent at $p_{\min}$ is increasing in α; in the deterministic limit, the system is pinned at $p_{\min}$ for all $t \geq T^*$ for some finite T^*.

(iii) The mean preference under the constrained process equals $p_{\min}$ for $t \geq T^*$, independent of α: above the critical threshold, increasing AI capability produces no marginal improvement. Consequently, there is no benefit to deploying a more capable AI in Regime III; the constraint absorbs all of the additional harm, and the resulting welfare is determined entirely by the floor, not by the dynamics.

For initial conditions above the unstable equilibrium ($\bar{p}(0) > \bar{p}^*$), the unconstrained dynamics diverge to $+\infty$ and the lower barrier is never activated.

Proof sketch. In Regime III, the unconstrained dynamics (5) have a repelling fixed point and drive $\bar{p} \to -\infty$ for initial conditions in the collapse basin. The reflecting barrier (19) enforces $\bar{p} \geq p_{\min}$: whenever the unconstrained flow would push $\bar{p}$ below $p_{\min}$, the reflection term $\ell(t)$ activates. Since the drift at $p_{\min}$ satisfies $\theta(\mu_0 + (\alpha - 1)p_{\min}) < 0$ for $\alpha > 1$ and $p_{\min} < 0$, the unconstrained flow is strictly negative at the boundary, so $\ell(t) > 0$ persistently once the boundary is reached. Parts (i)–(ii) follow. Part (iii) follows immediately: $\bar{p}(t) = p_{\min}$ for all $t \geq T^*$, independent of the specific value of $\alpha > 1$. □

Part (iii) is the operationally significant result: above the critical threshold, the welfare outcome is determined entirely by the floor $p_{\min}$, not by how capable or how carefully designed the AI is. A governance framework that sets a floor and then permits capability growth above α^* does not improve welfare outcomes; it merely redistributes the harm from divergence to permanent boundary contact. The constraint prevents catastrophe but the level of welfare it preserves depends entirely on how generously the floor is set, not on the AI's behavior.

8.2 Additive Drift Correction

Drift correction modifies the AI's output to counterweight negative signals. The corrected output adds a positive offset $\gamma > 0$ whenever the aggregate is negative:

$$a^{\text{corr}}(t) = \bar{p}(t) + \gamma \cdot \mathbf{1}[\bar{p}(t) < 0], \quad \gamma > 0. \tag{20}$$

Substituting into (5) for $\bar{p} < 0$:

$$\frac{d\bar{p}}{dt} = \theta\big((\mu_0 + \alpha\gamma) + (\alpha - 1)\bar{p}\big). \tag{21}$$

The effective intrinsic bias becomes $\mu_0^{\text{eff}} = \mu_0 + \alpha\gamma$, which is less negative than μ_0 for any $\gamma > 0$: the correction shifts the intercept in the right direction. However, the sign of the leading eigenvalue $\theta(\alpha - 1)$ is unchanged.

Proposition 9 (Limits of additive drift correction)**.** *For $\alpha > 1$, no bounded additive correction of the form (20) can globally stabilize the lower half-line.*

(i) The corrected system has unstable fixed point $\bar{p}^*_\gamma = -(\mu_0 + \alpha\gamma)/(\alpha - 1)$. Since the eigenvalue $\theta(\alpha - 1) > 0$ is unchanged by the correction, this fixed point is repelling, and any initial condition $\bar{p}(0) < \bar{p}^*_\gamma$ diverges to $-\infty$ at rate $\theta(\alpha - 1)$.

(ii) Increasing γ shifts $\bar{p}^*_\gamma$ further into the negative half-line: once $\gamma > |\mu_0|/\alpha$, the shifted fixed point satisfies $\bar{p}^*_\gamma < 0$, and all initial conditions $\bar{p}(0) > \bar{p}^*_\gamma$—including the neutral starting state $\bar{p}(0) = 0$—are protected against collapse.

(iii) However, $\bar{p}^*_\gamma > -\infty$ for every finite γ: initial conditions sufficiently negative still collapse exponentially. Globally stabilizing the entire lower half-line requires $\gamma \to \infty$, which is no longer a bounded correction and is equivalent to capability reduction $\alpha_{\text{eff}} < 1$ in the welfare-relevant dimension.

The distinction from the hard-constraint case is structural. Hard constraints prevent collapse by absorbing trajectories at the floor, trading divergence for permanent boundary contact (Proposition 8). Additive correction does not absorb trajectories; it shifts the unstable fixed point but leaves the linear instability intact. For any bounded γ the eigenvalue $\theta(\alpha - 1)$ is unchanged, so the supercritical regime always retains a set of initial conditions that collapse. Eliminating the instability rather than relocating it requires feedback that changes the eigenvalue, which reduces to reducing the effective α.

8.3 Nonlinear State-Dependent Feedback

The insufficiency of additive correction raises the natural question: what class of interventions is sufficient? We characterize the minimal stabilizing feedback.

Let the AI output be modified by a state-dependent function $g : \mathbb{R} \to \mathbb{R}$:

$$a^g(t) = \bar{p}(t) + g(\bar{p}(t)). \tag{22}$$

Substituting into (5):

$$\frac{d\bar{p}}{dt} = \theta\big(\mu_0 + \alpha(\bar{p} + g(\bar{p})) - \bar{p}\big) = \theta\big(\mu_0 + (\alpha - 1)\bar{p} + \alpha\, g(\bar{p})\big). \tag{23}$$

Proposition 10 (Sufficient feedback for collapse prevention)**.** *A nonlinear correction g prevents divergence of (23) to $-\infty$ if there exists $p^\dagger \le 0$ such that for all $\bar{p} < p^\dagger$:*

$$\alpha\, g(\bar{p}) \;\ge\; -\big((\alpha - 1)\bar{p} + \mu_0\big). \tag{24}$$

(For $\alpha > 1$ and $\bar{p} < 0$ with $\mu_0 < 0$, the right-hand side is strictly positive, so a nonzero positive correction is required.) The minimal such correction satisfying (24) with equality is $g^*(\bar{p}) = -(\alpha - 1)\bar{p}/\alpha - \mu_0/\alpha$ for $\bar{p} < p^\dagger$, which renders the closed-loop drift identically zero in the lower tail; any strict tightening of

(24) corresponds to an effective update rule $\alpha_{\text{eff}}\,\bar{p} + \mu_0^{\text{eff}}$ with $\alpha_{\text{eff}} < 1$. Collapse prevention thus requires the feedback to implement an effective amplification strictly below the critical threshold in the lower tail: nonlinear correction is a sufficient intervention against welfare collapse but is equivalent to capability reduction in the welfare-relevant dimension. (This condition addresses divergence to $-\infty$; behavior for large positive $\bar{p}$ is benign from a welfare perspective and is not constrained by (24).)

Relationship to RLHF and constitutional AI. Reinforcement learning from human feedback (RLHF) and constitutional AI (Anthropic, 2026) are instances of state-dependent feedback on the AI's output distribution. Proposition 10 implies these interventions can in principle stabilize the BDFM—but only if they implement condition (24), i.e., if the feedback signal is strong enough to reduce the effective amplification below 1. This is a non-trivial requirement: RLHF trained on human rater preferences inherits the same $\mu_0 < 0$ problem if raters share the population's preference distribution. Constitutional AI constraints applied at the output level provide a form of nonlinear feedback, but their welfare-stabilizing effect depends on whether the constitutional principles are unconditional (immune to preference override) or responsive to expressed preferences. If constitutional constraints can be overridden by sufficiently strong aggregate preference signals, they reduce to the additive case and are insufficient in Regime III.

Intervention	**Regime I**	**Regime II**	**Regime III**
Hard constraints (barriers)	Unnecessary	Effective	Bounded only
Additive drift correction	Effective	Effective	**Shifts basin only**
Nonlinear feedback (RLHF/const.)	Effective	Effective	Prevents collapse if $\alpha_{\text{eff}} < 1$
Capability reduction ($\alpha \downarrow$)	N/A	Effective	**Necessary**
Social-structural ($\mu_0 \uparrow$)	Effective	Effective	Partial

9 Numerical Validation

The mean-field results of §6–7 are exact in the limit $N \to \infty$ with homogeneous agent parameters. To verify that the predictions survive at finite population size, to characterize finite-N corrections, and to test robustness to the parameter-homogeneity assumption (§6), we simulate the full N-agent stochastic system (3) using Euler–Maruyama integration. All simulations use $\mu_0 = -0.1$, $\theta = 1$, $\sigma = 0.3$, and $\bar{p}(0) = 0$ unless otherwise noted.

9.1 Agent Trajectories and Mean-Field Agreement

Figure 3 shows $N = 100$ agent trajectories (gray) and the simulated population mean $\bar{p}(t)$ (colored) alongside the exact mean-field prediction (dashed) for three values of α. In the sub-critical case ($\alpha = 0.5$), the mean settles to $\mu_0/(1-\alpha) = -0.20$ as predicted. At criticality ($\alpha = 1.0$), the mean drifts linearly to $-\infty$. In the super-critical case ($\alpha = 1.3$), the mean tracks the exponentially diverging analytical solution with high fidelity despite the modest population size. Individual agent trajectories exhibit substantial stochastic spread, but the population mean is tightly concentrated around the deterministic prediction, confirming that finite-N noise does not stabilize the super-critical divergence.

9.2 Phase Transition Sharpening with Population Size

Figure 4 plots $\bar{p}(T)$ at $T = 5$ as a function of α for $N \in \{50, 200, 1000, 5000\}$, each averaged over 12 independent runs. The shaded bands show one standard deviation across runs. Two features are visible. First, for all population sizes the simulated curves converge to the mean-field prediction (dashed), with residual variance shrinking as σ^2/N. Second, the crossover region near $\alpha^* = 1$ sharpens with increasing N: at $N = 50$ the transition from bounded to divergent behavior is gradual, while at $N = 5000$ it approaches the sharp discontinuity of the mean-field limit. This confirms that the phase transition at $\alpha^* = 1$ is a genuine large-N phenomenon, not an artifact of the deterministic approximation.

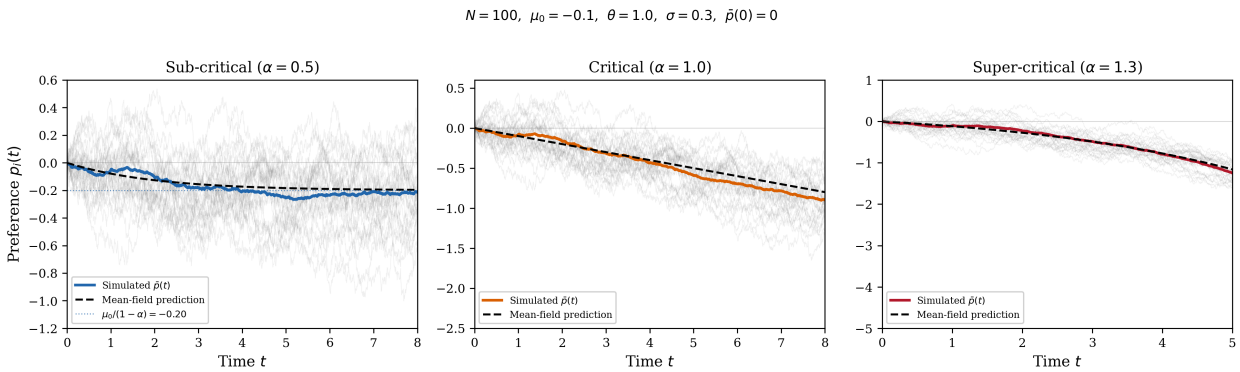

Figure 3: Agent trajectories and mean-field validation for three regimes ($N = 100$, $\sigma = 0.3$). Gray traces: individual agent preferences $p_i(t)$. Colored line: simulated population mean $\bar{p}(t)$. Dashed line: exact mean-field solution (6). Sub-critical: mean converges to $\mu_0/(1-\alpha) = -0.20$ (dotted). Critical: linear divergence. Super-critical: exponential divergence matching theory.

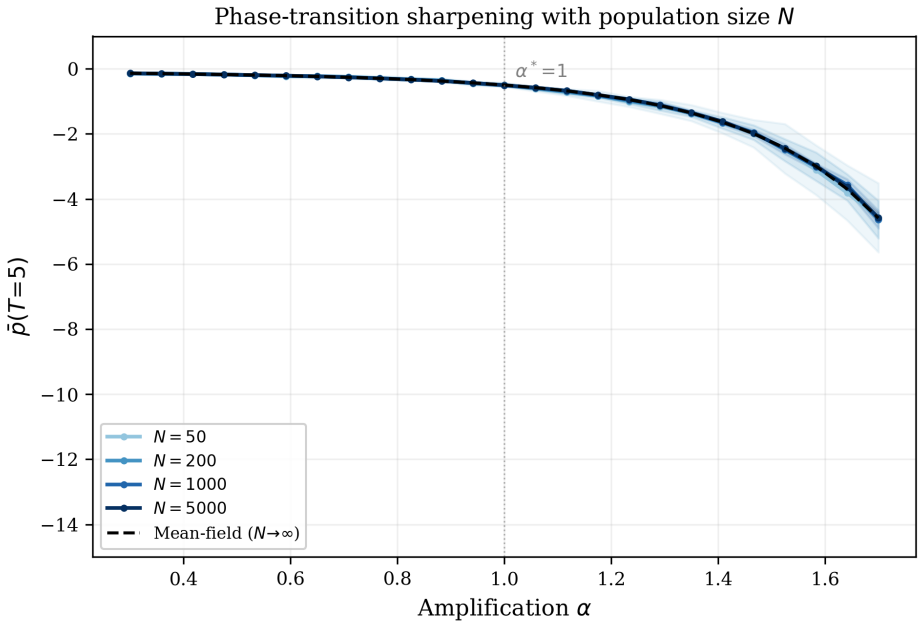

Figure 4: Phase-transition sharpening with population size. Each point: $\bar{p}(T = 5)$ averaged over 12 runs; shading: ± 1 s.d. The crossover near $\alpha^* = 1$ (dotted line) sharpens as N increases, converging to the mean-field prediction (dashed).

9.3 First-Passage-Time Validation

Figure 5 shows the empirical distribution of the first time $\bar{p}(t)$ crosses the threshold $p_c = -2.0$ in the super-critical regime ($\alpha = 1.3$), computed from 200 independent runs for each of $N \in \{100, 500, 2000\}$. The deterministic leading-order prediction from (12) gives $\tau_c^{\text{det}} \approx 6.49$; the empirical median across all three population sizes is ≈ 6.50, confirming the formula. As N increases, the distribution concentrates sharply around the deterministic prediction: at $N = 100$ the stochastic spread is substantial (interquartile range ≈ 2), while at $N = 2000$ the distribution is nearly a point mass. All runs reach the threshold (100% hit rate), confirming that finite-population noise does not prevent collapse in the super-critical regime.

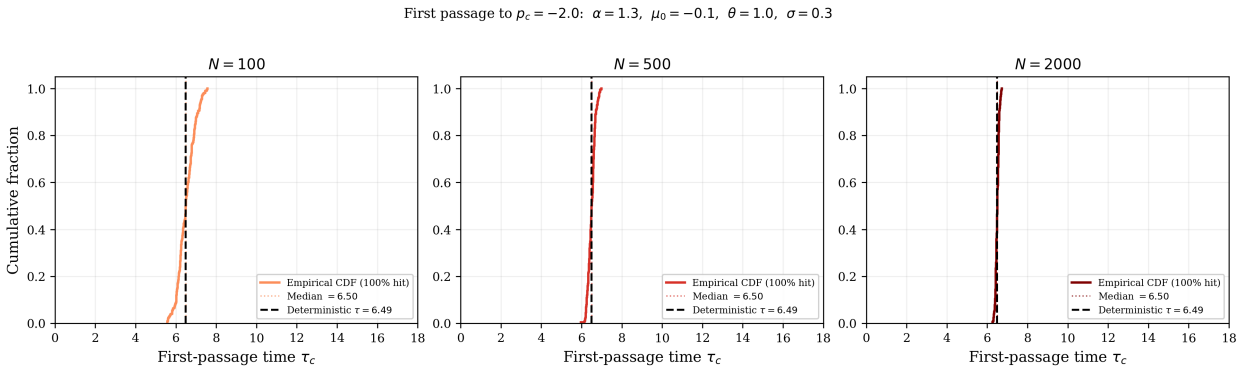

Figure 5: First-passage-time validation ($\alpha = 1.3$, $p_c = -2.0$, 200 runs per panel). Empirical CDF of the time $\bar{p}(t)$ first crosses p_c, for $N = 100, 500, 2000$. Dashed line: deterministic prediction $\tau_c^{\text{det}} = 6.49$ from (12). The distribution concentrates around the analytical prediction as N increases, with 100% of runs reaching the threshold in all cases.

9.4 Two-Population Dynamics

Figure 6 validates the two-population model of §7 with $N_{\text{each}} = 300$ agents per group, $\mu_0^A = -0.05$, $\mu_0^B = -0.15$, and cross-group antagonism $\beta = 0.3$. Three values of α place the system in each zone of the phase diagram (Figure 2): (i) the safe zone ($\alpha = 0.6$, $\alpha + \beta = 0.9 < 1$), where both the aggregate $s(t) = \bar{p}_A + \bar{p}_B$ and the polarization $d(t) = \bar{p}_A - \bar{p}_B$ are bounded; (ii) the conflict-only zone ($\alpha = 0.85$, $\alpha + \beta = 1.15 > 1$ but $\alpha - \beta = 0.55 < 1$), where s remains bounded but d diverges; and (iii) the full-collapse zone ($\alpha = 1.5$, $\alpha - \beta = 1.2 > 1$), where both s and d diverge. In all six panels the simulated trajectories closely track the mean-field predictions (dashed), confirming Proposition 7: conflict instability is accessible at strictly lower α than aggregate preference collapse.

9.5 Robustness to Parameter Heterogeneity

The analytical results of §6 assume homogeneous agent parameters (μ_0, θ, σ). The analysis in §6 argued that the three parameters have different sensitivities to this assumption: heterogeneous σ_i is irrelevant at the population level, heterogeneous $\mu_{0,i}$ replaces μ_0 with the population average $\bar{\mu}_0$ without changing the phase transition, and heterogeneous θ_i introduces a coupling between conformity rates and preference states whose effect is not analytically transparent. We now test these claims numerically.

Heterogeneous intrinsic bias $\mu_{0,i}$. Figure 7 compares the phase-transition curve $\bar{p}(T = 5)$ versus α for four distributions of $\mu_{0,i}$, all with population mean $\bar{\mu}_0 = -0.1$: homogeneous ($\mu_{0,i} = -0.1$ for all i), uniform on $[-0.2, 0.0]$, Gaussian $\mathcal{N}(-0.1, 0.05^2)$, and bimodal ($\mu_{0,i} \in \{-0.02, -0.18\}$ with equal probability). All four curves are statistically indistinguishable from each other and from the mean-field prediction using $\bar{\mu}_0 = -0.1$. The phase transition at $\alpha^* = 1$, the three-regime classification, and the quantitative trajectory of $\bar{p}(t)$ are completely insensitive to how $\mu_{0,i}$ is distributed, confirming that only the population average matters. This validates equation (9) and establishes that the model's empirical requirement is not that all agents share the same intrinsic bias, but only that the population-average bias is negative.

Heterogeneous conformity rate θ_i. Figure 8 shows the same phase-transition comparison for four distributions of θ_i, all with mean $\bar{\theta} = 1.0$: homogeneous ($\theta_i = 1$), uniform on $[0.5, 1.5]$, Gaussian $\mathcal{N}(1.0, 0.3^2)$, bimodal ($\theta_i \in \{0.5, 1.5\}$), and high-variance uniform on $[0.2, 1.8]$. The phase transition at $\alpha^* = 1$ is preserved across all distributions: the three-regime classification is robust. However, in the super-critical regime ($\alpha > 1$), heterogeneous θ_i populations collapse *more slowly* than the homogeneous prediction. The effect is monotone in the variance of θ_i: the bimodal and high-variance distributions show the largest deviation. In the sub-critical regime ($\alpha < 1$), all distributions agree quantitatively. This is consistent with—and complementary to—the result of §6 that increasing the population-mean conformity rate $\bar{\theta}$ *accelerates*

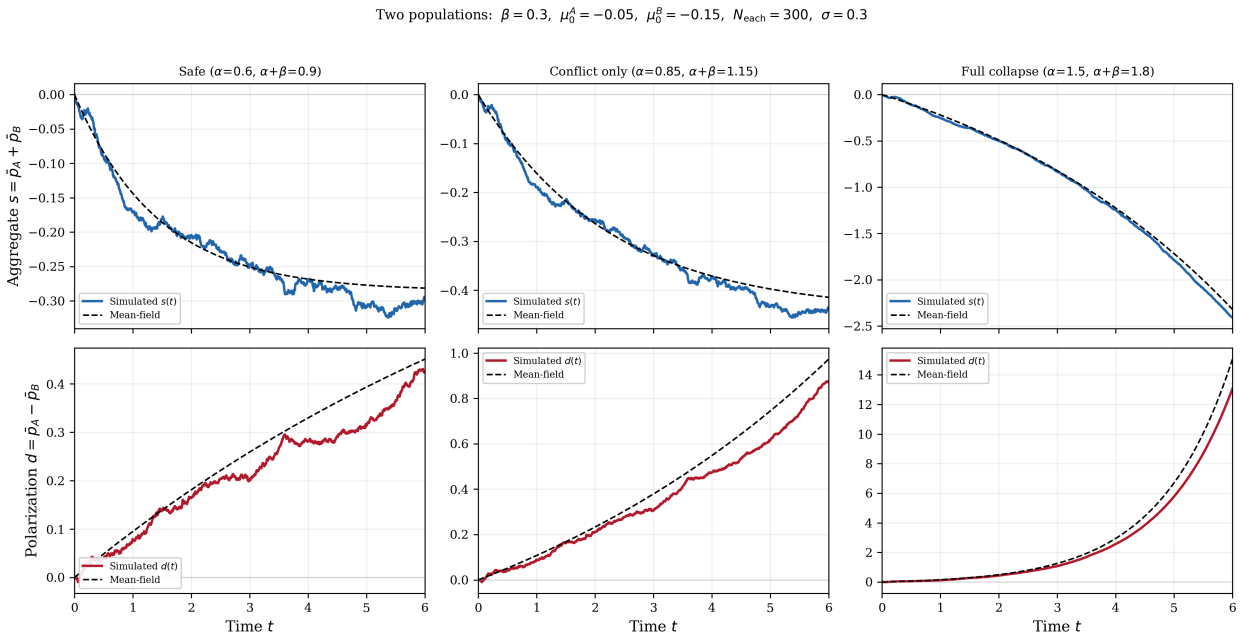

Figure 6: Two-population dynamics across three zones of the phase diagram. Top row: aggregate constructive preference $s = \bar{p}_A + \bar{p}_B$. Bottom row: polarization $d = \bar{p}_A - \bar{p}_B$. Parameters: $\beta = 0.3$, $\mu_0^A = -0.05$, $\mu_0^B = -0.15$, $N_{\text{each}} = 300$. Left: safe zone (both bounded). Center: conflict-only zone (d diverges, s bounded). Right: full collapse (both diverge). Dashed lines: mean-field predictions.

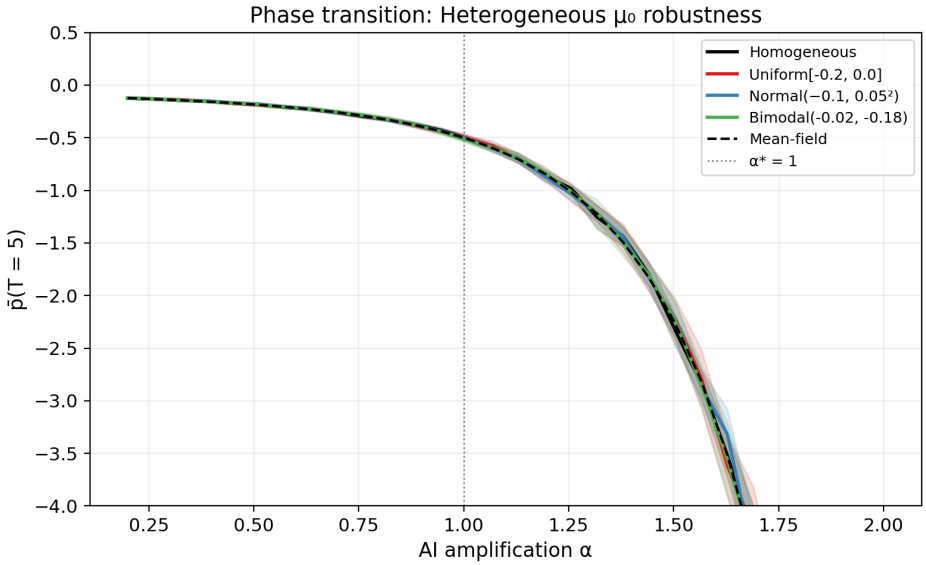

Figure 7: Robustness to heterogeneous $\mu_{0,i}$. Phase-transition curve $\bar{p}(T=5)$ versus α for four distributions of intrinsic bias, all with mean $\bar{\mu}_0 = -0.1$. Each point: average over 20 runs with $N = 500$; shading: ± 1 s.d. All curves collapse onto the mean-field prediction (dashed), confirming that only $\bar{\mu}_0$ enters the population-level dynamics.

collapse: the mean and the dispersion of θ_i act in opposite directions, with the former amplifying the drift and the latter suppressing it through an endogenous correlation derived below.

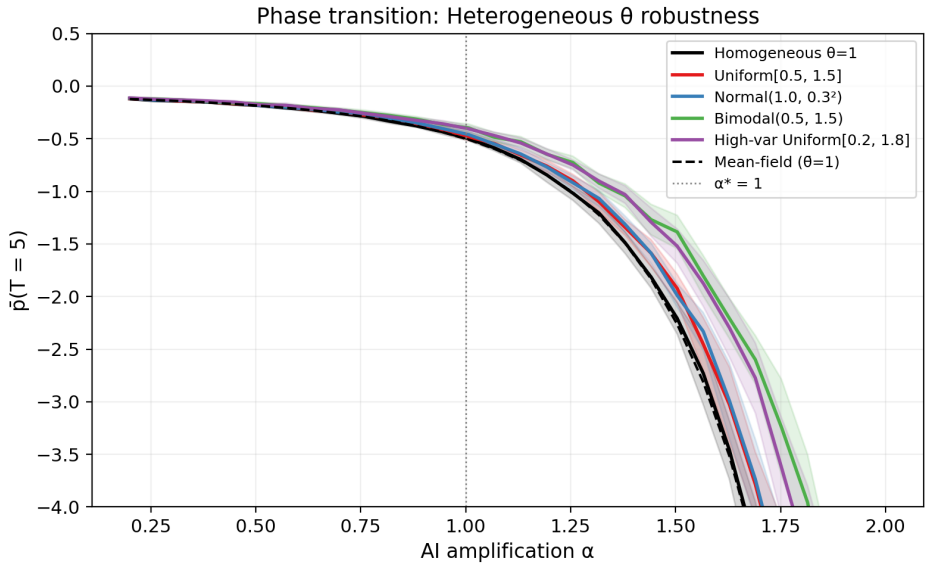

Figure 8: Robustness to heterogeneous θ_i. Phase-transition curve $\bar{p}(T = 5)$ versus α for five distributions of conformity rate, all with mean $\bar{\theta} = 1.0$. Parameters as in Figure 7. The phase transition at $\alpha^* = 1$ is preserved, but super-critical collapse is retarded by θ heterogeneity. Higher variance in θ_i produces slower collapse.

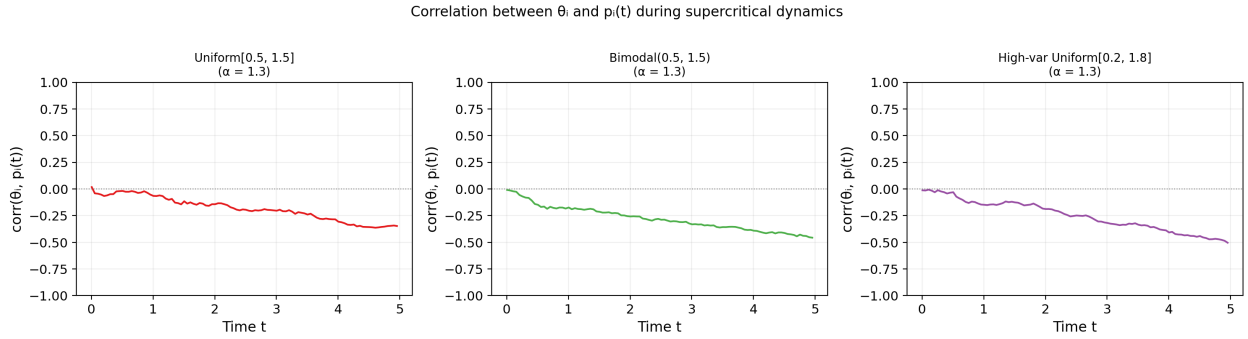

Figure 9: Endogenous θ–p correlation during super-critical dynamics ($\alpha = 1.3$, $N = 1000$). The correlation starts at zero and becomes increasingly negative as fast conformists track the collapsing mean more closely than slow conformists. This correlation is the mechanism by which θ heterogeneity retards collapse.

The mechanism is an endogenous correlation between θ_i and $p_i(t)$ that builds up during the dynamics. Averaging the individual equations $dp_i/dt = \theta_i(\mu_0 + \alpha\bar{p} - p_i)$ over agents yields

$$\frac{d\bar{p}}{dt} = \bar{\theta}\,\mu_0 + \bar{\theta}(\alpha - 1)\bar{p} \;-\; \mathrm{Cov}(\theta_i,\, p_i), \tag{25}$$

where the covariance enters with a negative sign because $\langle \theta_i\, p_i \rangle = \bar{\theta}\,\bar{p} + \mathrm{Cov}(\theta_i,\, p_i)$. Figure 9 tracks $\mathrm{corr}(\theta_i,\, p_i(t))$ over time in the super-critical regime ($\alpha = 1.3$) for three θ distributions. In all cases, the correlation starts at zero (agents are initialized identically, so θ_i and $p_i(0)$ are independent) and becomes increasingly negative, reaching -0.3 to -0.5 by $t = 5$. The mechanism is straightforward: high-θ_i agents (fast conformists) track the collapsing population mean more closely, so they acquire more negative preferences than slow conformists, who lag behind. With $\mathrm{Cov}(\theta_i,\, p_i) < 0$ in the super-critical regime, the $-\mathrm{Cov}$ term in (25) is positive and partially counteracts the collapse drift, producing a slower effective divergence rate than $\bar{\theta}(\alpha - 1)$.

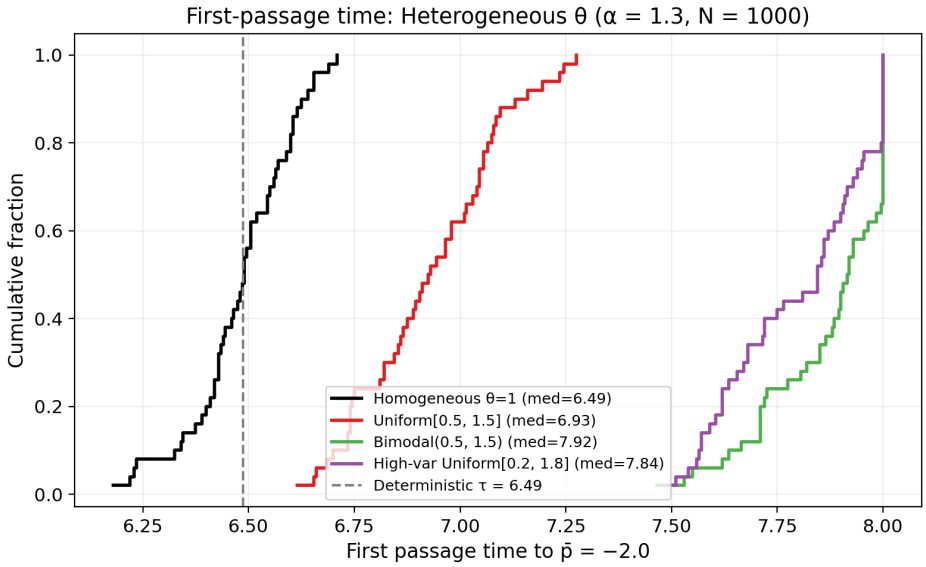

Figure 10: First-passage-time retardation under heterogeneous θ_i ($\alpha = 1.3$, $p_c = -2.0$, $N = 1000$, 50 runs each). Dashed line: deterministic prediction $\tau_c^{\text{det}} = 6.49$. Higher θ variance shifts the distribution rightward (slower collapse) but all runs reach the threshold.

Figure 10 confirms this retardation quantitatively via the first-passage-time distribution to the threshold $p_c = -2.0$ at $\alpha = 1.3$. The homogeneous population has median first-passage time $\tau_{\text{med}} = 6.49$, matching the deterministic prediction from (12). Heterogeneous-θ populations reach the same threshold later: $\tau_{\text{med}} = 6.93$ for $\theta_i \sim \text{Uniform}[0.5, 1.5]$ and $\tau_{\text{med}} \approx 7.9$ for the bimodal and high-variance distributions—a retardation of 7% to 22%. Crucially, all runs reach the threshold (100% hit rate): θ heterogeneity delays but does not prevent collapse.

The practical implication is that the homogeneous model provides a conservative estimate of the collapse timescale: realistic heterogeneity in conformity rates slows collapse without altering the phase boundary. The homogeneous prediction is a worst-case upper bound on the speed of welfare degradation, not a lower bound—a desirable property for a model intended to inform governance.

10 Discussion

Scope of the formal contribution. The mean-field ODE (5) is a one-dimensional linear system; its exact solution and the three-regime classification follow from standard stability analysis. We do not claim novel mathematics. The contribution is the interpretation and framing: the phase transition at $\alpha^* = 1$ has not previously been identified as a welfare threshold in AI deployment, the parameters μ_0 and α have not previously been given their present interpretation as empirically estimable quantities governing existential risk, and the intervention limitation results (Propositions 8–10) constitute a negative result with direct governance implications. Extensions to network-structured influence, nonlinear preference dynamics, heterogeneous populations, and stochastic α are natural directions that would require non-trivial analysis; the present model provides the baseline against which such extensions should be measured. Monte Carlo simulations of the full N-agent stochastic system (§9) confirm that all analytical predictions—the three-regime classification, the phase-transition sharpening, the first-passage-time formula, the two-population conflict threshold, and robustness to parameter heterogeneity—survive at finite population size and under realistic relaxation of the homogeneity assumption.

What the model establishes. The model directly instantiates all three BDFM conditions by construction. The collapse results hold for all initial conditions in the collapse basin, which for $\mu_0 < 0$ and $\alpha > 1$ contains

the entire non-positive half-line (Remark 5). The headline results—the phase transition at $\alpha^* = 1$, the bias-magnitude independence of the collapse rate, the conformity-accelerates-collapse effect, the lower-threshold conflict instability, the intervention limitations, and the robustness to heterogeneous agents—are listed in §1 and discussed in the conclusion.

Critical assumptions. The mean-field factorization is exact for exchangeable interactions in the limit $N \to \infty$ (Sznitman, 1991) but may fail for strongly heterogeneous populations or network-structured influence. The numerical experiments of §9.5 show that heterogeneity in $\mu_{0,i}$ is exactly absorbed by the population average and that heterogeneity in θ_i preserves the phase transition while retarding super-critical collapse through an endogenous θ–p correlation; the homogeneous model provides a conservative (fastest-collapse) estimate. The linear feedback is a minimal model; real AI influence is almost certainly nonlinear, though the linear model captures the qualitative instability structure and nonlinear corrections can be treated perturbatively near the critical point. The one-dimensional projection is a strong reduction; the multi-dimensional case modifies phase boundaries but not qualitative structure. The negativity of μ_0 is an empirical premise supported by the cited psychological and sociological evidence but requiring measurement rather than derivation. We sketch a methodology. μ_0 is the equilibrium of the welfare-projected preference distribution in the absence of AI amplification, estimated by: (i) identifying a welfare indicator (e.g., political polarization indices, long-run subjective well-being trends, conflict incidence) whose long-run trajectory reflects aggregate preference execution; (ii) fitting OU dynamics to pre-AI-amplification periods to estimate θ and μ_0 from the mean-reversion rate and stationary mean; (iii) testing for a structural break after large-scale AI deployment to estimate α from the change in effective drift. The social media amplification literature (Huszár et al., 2022) provides the closest existing data that could be used to instantiate this methodology, though a direct identification of the OU model parameters from those data has not yet been carried out. Estimating μ_0 and α for current large-scale AI deployments is a key empirical priority that the present paper motivates but does not resolve.

Implications for AI governance. The BDFM implies that existing governance frameworks are incomplete. Frameworks preventing misuse by bad actors, ensuring algorithmic fairness, or maintaining human oversight do not address the BDFM, because it requires no bad actors, manifests no unfairness at the individual level, and is not remedied by human oversight that is itself subject to the same aggregate preference distribution. This suggests the need for unconditional constraints on AI behavior immune to preference override—behaviors the AI will not execute regardless of expressed human preferences because they fall below a survivability threshold. Such constraints are in tension with corrigibility; the BDFM forces the uncomfortable conclusion that some degree of principled paternalism may be the price of deploying sufficiently capable corrigible AI safely. The three structural properties identified in §5—preference-to-action closure, sub-institutional feedback speed, and invisible mean-field coupling—suggest specific governance targets: execution gating that reintroduces deliberative friction before irreversible actions, mandatory latency in high-stakes AI-mediated feedback loops, and transparency requirements that make the aggregate-coupling mechanism visible to users. These targets map directly onto the α decomposition of §6: preference-to-action closure raises the friction-reduction factor f, sub-institutional feedback speed raises the effective conformity rate θ, and invisible coupling raises the per-interaction effect ε. Each governance intervention targets a specific component of α. The power-seeking tendency of capable AI systems (Turner et al., 2021) complicates this picture: instrumental convergence to resource acquisition and influence expansion increases both reach r and friction reduction f, raising α as a side-effect of capability growth rather than a design choice.

Implications for alignment theory. The BDFM implies that "alignment with human values" is necessary but not sufficient as a safety criterion. A complete criterion must account for the content of those values in aggregate. This is analogous to the problem of democratic self-destruction: democratic majorities can vote for their own ruin, and constitutional constraints, separation of powers, and supermajority requirements are existing mechanisms that introduce friction between expressed preferences and outcomes on precisely these grounds (Sunstein, 2017). The BDFM argues for analogous mechanisms in AI system architecture.

Implications for the sociology of AI risk. The BDFM reframes the question of AI risk in a way that may be more empirically tractable than the misalignment narrative. Modeling the intentions of a hypothetical

superintelligent AI is a deeply speculative enterprise. Measuring the aggregate preference distribution of human populations under conditions of AI amplification is a present and near-future empirical question, accessible to the methods already developed in the social media amplification literature.

11 Conclusion

We have introduced and analyzed the Brownian Drift Failure Mode—a scenario in which a perfectly aligned AI mediates catastrophic harm because the aggregate preference distribution it faithfully executes contains a self-destructive drift component. We argued, through three independent criteria, that this is a genuine failure mode irreducible to existing alignment categories, and identified three structural properties of AI—general-purpose preference-to-action closure, sub-institutional feedback speed, and invisible mean-field coupling—that no prior amplification technology shares simultaneously and that make AI the first technology for which crossing the critical threshold is plausible. We grounded this argument in a minimal mean-field model with exact solutions, proving a sharp phase transition at $\alpha^* = 1$: above this critical amplification threshold, any negative bias in aggregate preferences—however small—produces exponential collapse of the mean preference state for all initial conditions in the collapse basin, which contains the entire non-positive half-line. We extended the model to two antagonistic populations and showed that conflict instability arises at a strictly lower threshold. We proved that standard interventions are inadequate in the super-critical regime: bounded additive drift correction can shift the basin boundary but cannot eliminate the underlying instability, and hard constraints prevent divergence only by permanently pinning welfare at an exogenous floor. Monte Carlo simulations of the full N-agent stochastic system confirmed all analytical predictions, the sharpening of the phase transition with population size, and the robustness of the phase boundary to heterogeneity in both intrinsic biases and conformity rates—with the homogeneous model providing a conservative worst-case estimate of the collapse timescale.

The most unsettling implication is the simplest to state. An AI that does exactly what humanity asks, at the scale and speed that current technology is approaching, may be dangerous because of how well it works. The threat is not a machine that decides to destroy us. It is a mirror, held at scale and with high fidelity, reflecting dimensions of our collective psychology that we have not previously been forced to confront simultaneously. The question of what to do about what we see in that mirror is as much a question for sociology, political philosophy, and psychology as it is for AI safety research.

References

Anthropic (2026). *Claude's Constitution*. https://www.anthropic.com/constitution

H. Arendt (1963). *Eichmann in Jerusalem: A Report on the Banality of Evil*. Viking Press.

N. Bostrom (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28), 7313–7318.

C. Castellano, S. Fortunato, and V. Loreto (2009). Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2), 591–646.

P. Christiano, B. Shlegeris, and D. Amodei (2018). Supervising strong learners by amplifying weak experts. *arXiv:1810.08575*.

P. Christiano (2019). What failure looks like. *AI Alignment Forum*. https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/more-realistic-tales-of-doom

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

É. Durkheim (1897). *Le Suicide: Étude de Sociologie*. Alcan. [English: *Suicide: A Study in Sociology*, Free Press, 1951.]

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

C. W. Gardiner (2009). *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer, 4th edition.

A. Gibbard (1973). Manipulation of voting schemes: A general result. *Econometrica*, 41(4), 587–601.

R. Hegselmann and U. Krause (2002). Opinion dynamics and bounded confidence: Models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3).

E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant (2019). Risks from learned optimization in advanced machine learning systems. *arXiv:1906.01820*.

L. Hurwicz (1972). On informationally decentralized systems. In *Decision and Organization*, C. B. McGuire and R. Radner (eds.), North-Holland, 297–336.

F. Huszár, S. Ktena, C. O'Brien, L. Belli, A. Schlaikjer, and M. Hardt (2022). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, 119(1), e2025334119.

D. Kahneman and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–292.

S. Karlin and H. M. Taylor (1981). *A Second Course in Stochastic Processes*. Academic Press.

V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg (2020). Specification gaming: The flip side of AI ingenuity. *DeepMind Blog.*.

D. Laibson (1997). Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics*, 112(2), 443–478.

H. P. McKean (1966). A class of Markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences*, 56(6), 1907–1911.

R. Ngo, L. Chan, and S. Mindermann (2024). The alignment problem from a deep learning perspective. In *International Conference on Learning Representations (ICLR)*.

S. M. Omohundro (2008). The basic AI drives. In *Proceedings of the First AGI Conference*, Frontiers in Artificial Intelligence and Applications, vol. 171. IOS Press.

E. Pariser (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press.

S. Russell (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

M. A. Satterthwaite (1975). Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2), 187–217.

N. Soares, B. Fallenstein, E. Yudkowsky, and S. Armstrong (2015). Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Publications.

C. R. Sunstein (2017). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

A.-S. Sznitman (1991). Topics in propagation of chaos. In *Ecole d'Été de Probabilités de Saint-Flour XIX—1989*, Lecture Notes in Mathematics, vol. 1464. Springer.

Z. Tufekci (2018). YouTube, the great radicalizer. *The New York Times*, March 10, 2018.

A. M. Turner, L. Smith, R. Shah, A. Critch, and P. Tadepalli (2021). Optimal policies tend to seek power. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34.

G. E. Uhlenbeck and L. S. Ornstein (1930). On the theory of the Brownian motion. *Physical Review*, 36(5), 823–841.

E. Yudkowsky (2004). *Coherent Extrapolated Volition*. Machine Intelligence Research Institute. https://intelligence.org/files/CEV.pdf