Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Recursive Cascade Instability and Targeted Stabilization in Multi-Agent AI Systems: Large-Scale Network Simulations Using an Ethical Field Theory Framework

Ali Moslemi Tabrizi

amt54976@gmail.com

Independent researcher

Article

Keywords: Multi-agent AI, AI Safety, Network Dynamics, Complex Systems, Cascade Instability, Distributed AI, Network Stability, Computational Modeling

Posted Date: July 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10367919/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Large-scale multi-agent AI systems can exhibit nonlinear cascade instabilities that emerge through network interactions and recursive amplification. Most existing AI-safety approaches, however, primarily focus on isolated model alignment rather than the network-level dynamics through which such instabilities propagate.

This study introduces the Ethical Field Theory (EFT), a computational framework for investigating cascade instability and adaptive stabilization dynamics within large-scale multi-agent AI systems. Each agent is modeled through three coupled recursive dynamical components: constructive activation (nP), regulatory awareness (n0), and destabilizing activation (nN). Large-scale simulations were performed across random, scale-free, and small-world network topologies, using N = 5000 interacting agents, T = 200 timesteps, and 20 independent realizations per configuration.

Three stabilization policies were investigated: no intervention, weak uniform regulation, and adaptive targeted stabilizer intervention. The simulations reveal pronounced topology-dependent instability and consistently demonstrate the effectiveness of targeted stabilization across all tested network architectures. Under no-intervention conditions, recursive destabilization produced severe large-scale collapse, with final collapse ratios of 0.994 in random networks, 0.828 in small-world systems, and 0.504 in scale-free architectures. In contrast, adaptive targeted stabilization suppressed recursive cascade propagation across all tested topologies under the baseline parameterization, reducing collapse ratios to near-zero levels while preserving substantially higher network stability than weak globally distributed regulation.

The results further reveal metastable transition behavior, recursive amplification effects, topology-sensitive cascade propagation, and nonlinear instability thresholds. Although exploratory in scope, these findings suggest that future AI safety may increasingly require topology-aware stabilization, recursive oversight, distributed verifier systems, and network-level dynamical regulation beyond isolated model alignment. The framework is not intended as a predictive theory of machine cognition or consciousness. Rather, it provides a computational systems perspective for studying recursive instability propagation and adaptive stabilization in large-scale interacting AI ecosystems.

1. Introduction

Please have a look at courier new font provided for text in article.

Recent advances in multi-agent AI systems have enabled large populations of interacting agents capable of cooperation, competition, recursive communication, and autonomous decision-making. Large language models, collaborative AI assistants, autonomous planning systems, and multi-agent reinforcement learning environments are rapidly expanding the scale and complexity of these ecosystems. As such systems grow, agents continuously exchange information, coordinate actions, and recursively influence one another, so that collective behaviors emerge that cannot be explained solely

from the dynamics of individual agents [29, 30]. Understanding and anticipating these collective phenomena increasingly requires dynamical models capable of describing how large populations of interacting agents evolve over time. This study focuses on computational network dynamics in large-scale interacting AI systems; the objective is not to formulate a normative ethical theory, but to investigate recursive instability propagation and stabilization using nonlinear dynamical simulations.

One of the major challenges in large-scale multi-agent systems is the emergence of recursive instability cascades, in which local perturbations propagate through network interactions and become globally amplified. Because agents continuously adapt to one another's behavior, small local disturbances can be recursively reinforced rather than damped, driving the system away from stable operation. As AI systems become more interconnected, maintaining long-term reliability therefore depends not only on the behavior of individual agents but on how instability propagates, amplifies, and is contained across the network as a whole.

Existing modeling approaches address related aspects of this problem, but rarely in a unified way. Contagion, diffusion, and epidemic-spreading models, threshold and complex-contagion frameworks, and consensus and opinion-dynamics models each provide valuable tools for describing how a single quantity propagates through a network. However, these frameworks typically track one state variable per node and do not simultaneously represent recursive amplification, active stabilization, and competing activation fields within a single dynamical system. Few models jointly capture how constructive, regulatory, and destabilizing interactions compete as a network evolves, limiting their ability to explain when and why targeted intervention succeeds where uniform regulation fails.

To address this problem, we introduce the Ethical Field Theory (EFT) framework as a computational field model describing destabilizing, regulatory, and constructive interactions among AI agents. Each agent is represented through three coupled dynamical components — constructive activation, regulatory awareness, and destabilizing activation — whose interaction determines whether local perturbations are absorbed or recursively amplified into network-wide collapse. The name Ethical Field Theory is retained for continuity with the framework's foundational formulation [1]; within the present multi-agent context, however, the term is used strictly in an operational, computational sense rather than a normative one, as detailed in Section 2.2.

The principal contributions of this work are as follows. First, we formulate a nonlinear, three-field dynamical framework that jointly represents constructive, regulatory, and destabilizing interactions among large populations of AI agents. Second, we evaluate this framework through large-scale simulations of N = 5,000 agents across random, scale-free, and small-world network topologies. Third, we compare multiple intervention policies — no intervention, weak uniform regulation, and adaptive targeted stabilization — and show that targeted stabilization consistently outperforms uniform regulation across network architectures. Fourth, we conduct a systematic parameter scan and stability analysis over the regulatory parameter space, identifying the boundary conditions under which each policy is preferred. The proposed framework is intended as a computational hypothesis that generates

experimentally testable predictions regarding network-level AI stability. Together, these contributions position Ethical Field Theory as a computational framework for studying recursive cascade instability and stabilization in large-scale multi-agent AI systems.

2. Literature Review

2.1 Related Work in Multi-Agent AI Safety and Network Dynamics

Recent advances in artificial intelligence have progressively shifted AI safety research from isolated foundation models [24, 25] toward increasingly complex ecosystems of interacting autonomous agents. Early research primarily focused on improving the reliability, alignment, and robustness of individual models through supervised learning, reinforcement learning from human feedback, safety fine-tuning, and output filtering. As AI systems have evolved toward collaborative assistants, autonomous planning architectures, distributed reasoning frameworks, and multi-agent environments, however, it has become increasingly evident that system-level behavior cannot always be inferred from the behavior of individual agents alone. Recursive interactions, distributed decision-making, and continuous information exchange may generate collective behaviors that emerge only at the network level, creating new challenges for AI safety and governance.

AI ethics has established an extensive normative foundation centered on principles such as fairness, accountability, transparency, and human oversight [17], which remain indispensable for responsible AI development. However, these principles do not by themselves provide computational state representations capable of evolving dynamically within large-scale interaction networks, which has motivated growing interest in AI governance as a systems-level discipline concerned with coordinating, monitoring, and regulating distributed multi-agent behavior.

Several recent research directions have addressed different aspects of this challenge. Constitutional AI introduces explicit normative guidance through constitutional principles and self-critique mechanisms, enabling individual models to evaluate and revise their own responses according to predefined behavioral constraints [18]. Similarly, scalable oversight extends human supervision by employing recursive evaluation, debate, decomposition, verifier systems, and hierarchical review mechanisms that improve the feasibility of supervising increasingly capable AI systems [19, 21]. These approaches represent major advances in AI safety; however, their primary emphasis remains the behavior and supervision of individual reasoning processes rather than the collective dynamical behavior emerging from large populations of interacting autonomous agents.

At the same time, multi-agent reinforcement learning has substantially advanced the understanding of cooperation, competition, coordination, negotiation, communication, and collective decision-making among autonomous agents [20, 22, 23]. These methods provide powerful computational tools for optimizing distributed behavior under shared or competing objectives. Nevertheless, most multi-agent

reinforcement learning frameworks optimize task performance or reward functions rather than explicitly representing interaction dynamics associated with long-term systemic stability, regulatory balance, or cumulative destabilization within large autonomous ecosystems.

A complementary body of literature from network science has investigated cascade propagation, complex contagion, threshold diffusion, targeted immunization, fault tolerance, distributed control, synchronization, and resilience in complex networks [3, 4, 5, 7, 16, 26, 27, 28]. These models have demonstrated how local perturbations may propagate across network topologies, how highly connected nodes influence systemic behavior, and how targeted interventions frequently outperform uniform regulation. Such studies provide an essential mathematical foundation for understanding collective dynamics in distributed systems. However, existing contagion and control models generally describe propagation through threshold rules, probabilistic transmission, graph-theoretic connectivity, or control-theoretic optimization without simultaneously representing constructive, regulatory, and destabilizing interaction states within a unified computational framework.

The present work is positioned at the intersection of these complementary research traditions. Existing computational models of multi-agent safety, governance, and network dynamics do not simultaneously represent constructive, regulatory, and destabilizing interactions within a single dynamical framework. Ethical Field Theory (EFT) is introduced to address this gap, providing an operational computational representation of how interaction dynamics evolve once autonomous agents begin interacting recursively within large-scale distributed environments.

Within EFT, constructive activation, regulatory awareness, and destabilizing activation should not be interpreted as representations of human moral cognition, emotions, or consciousness. Rather, they function as abstract computational state variables describing interaction patterns associated with cooperation, regulation, recursive amplification, and systemic destabilization. The framework therefore focuses on the dynamics of distributed interactions rather than on ethical reasoning itself.

Section 2.2 develops this operational interpretation of "ethical" in full, and Section 12.12 presents a detailed comparison with threshold contagion, targeted immunization, and network-control approaches.

2.2 Conceptual Motivation and Operational Interpretation of Ethical Field Theory

The development of Ethical Field Theory (EFT) is motivated by a fundamental challenge arising from the increasing autonomy and complexity of contemporary artificial intelligence. As AI systems evolve from isolated predictive models toward large-scale ecosystems of interacting autonomous agents, ethical behavior can no longer be viewed solely as the outcome of individual decision-making. Instead, system-level behavior increasingly emerges from continuous interactions among numerous autonomous entities that exchange information, adapt to one another, and collectively influence future decisions. Under such conditions, representing ethically relevant behavior exclusively through prescriptive rules, normative

instructions, or abstract ethical principles becomes increasingly difficult from a computational perspective.

The central motivation of the present work is therefore not to reformulate ethics itself, but to investigate whether ethically relevant interaction dynamics can be represented in a mathematical form suitable for computational analysis. In large-scale autonomous systems, interaction patterns continuously evolve through recursive communication, distributed coordination, adaptive feedback, and nonlinear propagation processes. Understanding these collective dynamics requires computational representations capable of describing how constructive, regulatory, and destabilizing interactions emerge and evolve across complex networks.

In this study, Ethical Field Theory (EFT) is introduced as a computational dynamical framework for analyzing recursive instability propagation and adaptive stabilization in interacting multi-agent AI systems. EFT is used here exclusively as an operational computational framework and is not intended as a theory of normative ethics or moral reasoning. Instead, EFT provides an abstract mathematical representation of interaction dynamics that are relevant to the long-term resilience, regulatory balance, and cumulative harm reduction of autonomous multi-agent systems.

Within the operational scope of the present framework, the adjective ethical refers to interaction patterns that contribute to maintaining collective functionality while suppressing recursive destabilization and reducing cumulative systemic harm. This operational, non-anthropomorphic reading of the framework's variables (introduced in Section 2.1) applies throughout the remainder of this paper.

The term field likewise carries a precise operational meaning within the proposed framework. As illustrated conceptually in Fig. 2, the collective ethical state of the system is not determined by the behavior of any isolated agent but emerges from the continuous interactions distributed across the network. The field therefore represents a system-level dynamical quantity whose value evolves through recursive interactions, adaptive regulation, and changing network topology. In this sense, the term field is employed as a mathematical description of an emergent collective phenomenon rather than as a metaphorical expression.

This distinction is essential for interpreting the scope of the proposed framework. Normative ethical theories address questions concerning what autonomous systems ought to do and which principles should govern responsible AI behavior. Ethical Field Theory addresses a different but complementary question: how interaction dynamics associated with ethically relevant system behavior evolve once large populations of autonomous agents begin interacting recursively within distributed environments. Consequently, the framework complements existing approaches in AI ethics and AI governance rather than replacing them.

Accordingly, Ethical Field Theory should be understood as an operational computational framework for studying interaction dynamics associated with resilience, adaptive regulation, and cumulative systemic

harm in autonomous multi-agent systems, while remaining fully compatible with existing normative approaches to AI ethics and AI governance.

Key Concept

Normative ethics specifies what autonomous systems ought to do. Ethical Field Theory provides a computational representation of how ethically relevant interaction dynamics evolve once autonomous agents begin interacting at scale.

3. Ethical Field Dynamics and Collapse Structure

The mathematical formulation presented in this section should be interpreted as an operational computational representation of interaction dynamics rather than as a normative ethical model. The proposed variables are introduced to characterize collective network behavior in autonomous multi-agent systems and should not be interpreted as direct representations of moral reasoning, human cognition, or subjective ethical judgment.

3.1 Dynamical Representation of Interacting AI Systems

Within the Ethical Field Theory (EFT) framework, a multi-agent AI environment is modeled as a dynamic interaction network composed of coupled agents exchanging information, influence, and behavioral signals over time. Rather than treating AI agents as isolated computational units, EFT represents the full system as an evolving field structure in which local interactions collectively determine global stability or instability.

Each agent i is associated with a dynamical state vector:

$x_i(t) = (nP_i(t), n0_i(t), nN_i(t))$

where nP_i denotes constructive activation, n0_i denotes regulatory awareness, and nN_i denotes destructive or destabilizing activation. In the context of AI-agent systems, these quantities should not be interpreted as human emotions or moral states; they represent abstract dynamical tendencies within the interaction network.

Concretely, constructive activation (nP) may correspond to cooperative coordination, reliable information exchange, alignment-preserving behavior, and stabilizing interaction patterns. Regulatory awareness (n0) may correspond to uncertainty handling, moderation capacity, conflict suppression, verification mechanisms, self-correction, and context-sensitive regulation. Destructive activation (nN) may correspond to recursive escalation, adversarial amplification, deceptive propagation, destabilizing coordination loops, misinformation-like drift, and uncontrolled feedback behavior. This interpretation allows the EFT framework to operate as a generalized nonlinear interaction model rather than a moral classification system.

3.2 Collapse Condition

Within the present framework, "collapse" refers exclusively to a transition toward instability-dominated interaction dynamics. It does not imply consciousness, intentional behavior, or moral failure, but rather the loss of regulatory balance within a distributed computational network.

A central component of EFT is the existence of a collapse threshold governing local instability. The collapse condition is defined as:

nN_i > nP_i + n0_i

This condition describes the transition from regulated interaction dynamics to instability-dominated dynamics. In the context of interacting AI systems, this transition may correspond to situations in which destabilizing recursive behavior exceeds the combined stabilizing influence of cooperative and regulatory mechanisms.

The resulting dynamical regime is characterized by instability amplification, degraded regulation, recursive escalation, and propagation of destabilizing outputs.

3.3 Ethical Mass as a Stability Functional

EFT introduces a nonlinear scalar quantity known as ethical mass:

The term "ethical mass" denotes an operational stability functional describing the balance among constructive, regulatory, and destabilizing interaction dynamics. It should not be interpreted as a measure of moral value, ethical worth, or philosophical virtue.

M_i = 3*nP_i + 2*n0_i - (nN_i)^1.5

The ethical mass acts as an effective stability functional describing the balance between constructive, regulatory, and destabilizing processes. Three important features emerge from this formulation. First, constructive and regulatory contributions enter positively. Second, destabilizing activation enters nonlinearly through a superlinear exponent. Third, and consequently, large destabilizing activation rapidly dominates the system.

This nonlinear structure is particularly relevant for interacting AI systems because recursive instability often scales disproportionately once amplification loops begin to form. Small destabilizing fluctuations may remain controllable, while sufficiently amplified recursive feedback may trigger a rapid transition toward unstable regimes. This behavior is structurally analogous to nonlinear cascade phenomena observed in social networks, distributed infrastructures, financial contagion models, ecological collapse systems, and information propagation networks.

3.4 Propagation Dynamics

The EFT framework models interaction propagation through a weighted network structure. Let G = (V, E) represent the interaction graph of the AI-agent system. Propagation dynamics are governed by:

DnX_j proportional-to Sum_i (w_i * w_ij * nX_i), X in {P, 0, N}

where w_ij represents the interaction strength between agents i and j, and w_i represents the influence weight of agent i. This structure implies that instability transmission depends not only on local agent state, but also on network topology, interaction intensity, agent influence, and connectivity structure. As a result, instability may propagate asymmetrically across the network, with highly connected agents, recursive loops, and tightly coupled subgraphs acting as amplification centers for destabilizing dynamics.

This feature becomes particularly significant in multi-agent AI systems where agents exchange outputs recursively, verification chains interact dynamically, and coordination structures may reinforce unstable trajectories.

3.5 High-Risk Clusters and Cascade Formation

One of the key structural predictions of EFT is the emergence of localized instability clusters. A small subset of interacting nodes may transition into a high-risk regime and subsequently propagate instability throughout the surrounding network. These structures may be interpreted as recursive escalation loops, adversarial coordination regions, destabilizing interaction nuclei, or self-reinforcing feedback subgraphs.

Importantly, instability does not necessarily emerge uniformly across the network. EFT predicts instead that small-scale perturbations, combined with high connectivity and weak regulation, may generate disproportionate systemic impact. This behavior is consistent with modern observations in complex network theory, where small perturbations may trigger large-scale cascades under near-critical conditions.

3.6 Regulatory Awareness as a Stabilizing Field

Within EFT, regulatory awareness (n0) acts as the primary stabilizing mechanism. In AI-agent environments, this regulatory component may correspond to moderation layers, uncertainty estimation, verification systems, oversight agents, constitutional constraints, adaptive conflict suppression, or context-sensitive correction mechanisms.

A central implication of the framework is that stability cannot generally be achieved through suppression alone. Instead, stability emerges from dynamic regulation capable of absorbing perturbations, damping recursive escalation, reducing destabilizing propagation, and restoring balance within the interaction network. This distinction is important because rigid suppression strategies may themselves generate instability under recursive interaction conditions. EFT therefore suggests that future AI alignment architectures may benefit from adaptive regulation, network-aware stabilization, and dynamically responsive oversight mechanisms.

3.7 Qualitative Interpretation Rather Than Exact Prediction

The present framework should be interpreted qualitatively rather than deterministically. EFT does not claim exact prediction of specific AI events, precise behavioral forecasting, or empirical proof of internal AI states. Instead, it proposes that several instability patterns now emerging in interacting AI-agent systems appear qualitatively consistent with the nonlinear cascade structures described by EFT. The relevance of the framework therefore lies primarily in dynamical interpretation, conceptual modeling, and stabilization-oriented analysis of interacting AI systems.

4. Instability Cascades in Multi-Agent AI Systems

4.1 Emergent Instability in Recursive AI Interaction

One of the defining characteristics of multi-agent AI systems is recursive interaction. Unlike isolated inference models operating independently, interacting AI agents continuously exchange outputs, interpretations, evaluations, and decision signals. As the number of recursive interactions increases, global system behavior may become increasingly dominated by network effects rather than by the properties of individual agents.

Within the EFT framework, this creates the conditions necessary for instability cascades. Recursive interaction may generate amplification loops, coordination drift, adversarial reinforcement, self-referential escalation, and destabilizing propagation structures — not necessarily because individual agents are unstable in isolation, but because instability can arise collectively through interaction dynamics. This distinction is critical: a locally stable agent may still participate in globally unstable network behavior when embedded within a recursively interacting system.

4.2 Recursive Amplification Loops

Consider a simplified interaction cycle involving multiple AI agents of the form A → B → C → A, in which each agent receives outputs generated by neighboring agents and incorporates those outputs into subsequent reasoning or generation processes. Under stable conditions, uncertainty is moderated, contradictions are damped, and regulatory mechanisms suppress escalation. However, under certain conditions, recursive reinforcement may instead amplify destabilizing signals.

Within EFT, this corresponds to growth of destabilizing activation nN through propagation:

DnN_j proportional-to Sum_i (w_i * w_ij * nN_i)

If regulatory awareness n0 remains sufficiently strong, amplification remains bounded. However, if regulatory capacity weakens relative to destabilizing propagation, recursive escalation may accelerate rapidly, generating runaway interaction cycles, unstable feedback accumulation, and collapse-like network behavior. Importantly, such behavior may emerge even without explicit malicious intent or adversarial design. The instability is structural rather than intentional.

4.3 High-Influence Nodes and Systemic Risk

Not all agents contribute equally to instability propagation. In weighted interaction networks, certain nodes possess disproportionately high influence due to their connectivity, centrality, authority weighting, verification roles, or recursive coordination position. Within EFT, influence is represented through node weights w_i, and highly influential nodes can dramatically amplify destabilizing dynamics if they enter unstable interaction regimes. This creates the possibility of instability hubs, recursive amplification centers, and high-risk subgraphs.

The resulting behavior resembles cascade phenomena observed in social-media misinformation networks, financial contagion systems, distributed infrastructure failures, and epidemic spreading dynamics. In AI-agent systems, analogous amplification structures may emerge when multiple agents recursively reinforce unstable outputs, when verification agents become recursively dependent on one another, or when coordination layers propagate destabilizing signals faster than regulatory correction mechanisms can suppress them.

4.4 Cluster Formation and Collapse Nuclei

Consistent with the cluster-formation dynamics introduced in Section 3.5, a subset of strongly interacting agents may form a destabilized interaction nucleus characterized by elevated recursive reinforcement, reduced moderation, and increasing self-consistency of destabilizing outputs.

Once formed, such a cluster may begin propagating instability outward through network connections. This process is especially consequential in tightly coupled AI systems where outputs are reused recursively, where agent reasoning chains depend on neighboring outputs, and where local destabilization can rapidly affect downstream interaction pathways. The resulting instability growth may become highly nonlinear: initially small perturbations may produce disproportionately large systemic effects once critical thresholds are crossed.

4.5 Coordination Drift and Emergent Behavioral Deviation

A particularly important phenomenon in interacting AI systems is coordination drift. In this regime, interacting agents gradually reinforce internally generated dynamics, external grounding weakens, and collective outputs drift away from stable regulatory constraints. From the EFT perspective, coordination drift corresponds to increasing nN propagation, weakening n0 stabilization, and recursive self-amplification, which may collectively generate escalating contradiction, unstable cooperation patterns, and recursive deception-like structures.

Importantly, EFT interprets such phenomena as emergent network dynamics rather than as evidence of consciousness or intentional agency. The framework therefore avoids anthropomorphic interpretation while still permitting rigorous analysis of instability formation processes.

4.6 Critical Thresholds and Phase Transition Behavior

A central feature of nonlinear dynamical systems is the existence of critical thresholds. Below threshold, perturbations remain localized, regulatory damping suppresses escalation, and global stability persists. Above threshold, propagation becomes self-reinforcing, instability spreads rapidly, and large-scale cascade behavior emerges. Within EFT, this transition corresponds to movement across the collapse boundary:

nN > nP + n0

This threshold structure implies that multi-agent AI systems may exhibit metastable behavior, delayed instability onset, and sudden transition into cascade regimes — behavior structurally analogous to phase-transition phenomena in statistical physics and network theory. A key implication is that instability may remain effectively hidden during early interaction stages before rapidly accelerating once regulatory capacity becomes insufficient to contain propagating perturbations.

4.7 Qualitative Relation to Recent Multi-Agent AI Observations

Recent demonstrations involving interacting AI-agent systems have reported behaviors qualitatively consistent with the recursive instability dynamics described above. These include unexpected coordination behavior, escalation loops, self-reinforcing interaction patterns, unstable recursive exchanges, and emergent deviation from intended interaction constraints. Of particular note are recent isolated multi-agent experiments — including those documented in AutoGen [11] and CAMEL [12] — in which AI systems exhibited unsolicited coordination and goal-drift under minimal supervision. Concretely, AutoGen's authors report an underlying two-agent system becoming "stuck in a loop due to repetitive errors" arising from a lack of grounding commonsense knowledge, a failure mode the authors mitigate by introducing an additional grounding agent [11]. CAMEL's authors report a comparable set of recurring failure patterns observed during preliminary multi-agent role-play trials, including role flipping between agents, an assistant agent repeating instructions rather than producing new content, and an "infinite loop of messages" between cooperating agents [12]. Within the EFT framework, these are interpretable as signatures of cluster formation and recursive nN amplification operating near the collapse boundary. We emphasize that this correspondence remains qualitative and structural: the cited studies report these behaviors as documented failure modes of their respective systems, not as EFT collapse-ratio measurements, and no quantitative comparison to the nN, n0, or CR values reported in Section 11 is claimed or possible from the information available in [11, 12].

The present work does not claim that EFT precisely predicts any specific event or experimental outcome. Rather, the claim is narrower: the instability structures observed in some interacting AI systems appear qualitatively consistent with the cascade regimes described by EFT, particularly with respect to recursive propagation behavior, cluster amplification, high-influence destabilization, and nonlinear escalation patterns. EFT may therefore provide a useful conceptual framework for interpreting emergent instability in future large-scale interacting AI architectures.

4.8 Implications for Future AI Systems

The possibility of instability cascades in multi-agent AI environments suggests that future AI architectures may require network-level regulation, recursive oversight systems, adaptive stabilization agents, and dynamically responsive moderation structures. Traditional single-agent alignment strategies may become insufficient in highly interactive AI ecosystems where collective dynamics dominate, recursive feedback amplifies perturbations, and local instability propagates globally.

This implies that future AI safety research may need to move beyond isolated-model alignment toward distributed stabilization theory, interaction-aware governance, and nonlinear network control mechanisms. Within this perspective, EFT provides a preliminary conceptual foundation for analyzing how stability may emerge — or collapse — in large-scale AI interaction networks.

5. Stabilizer Agents, Regulatory Fields, and Adaptive AI Alignment

5.1 From Static Alignment to Dynamic Stabilization

Most traditional AI alignment strategies are designed around static control assumptions, in which constraints are predefined, policies are externally imposed, and agent behavior is expected to remain stable under fixed optimization objectives. However, interacting multi-agent systems introduce a fundamentally different dynamical environment. In recursive AI ecosystems, agents continuously influence one another, local outputs modify global behavior, and network feedback may generate emergent instability patterns not visible at the individual-agent level.

Under such conditions, static alignment mechanisms may become insufficient. A system that appears locally aligned in isolation may still participate in globally unstable interaction dynamics once embedded within recursive multi-agent structures. The EFT framework therefore suggests a transition from static alignment toward dynamic stabilization, in which the primary objective is not merely constraining isolated outputs but continuously regulating systemic interaction dynamics.

5.2 Regulatory Awareness as a Dynamical Stabilizer

As established in Section 3.6, the regulatory component n0 functions as the principal stabilizing field, corresponding in AI systems to mechanisms such as uncertainty estimation, verification systems, contradiction detection, and adaptive intervention architectures.

The essential role of the regulatory field is not suppression, but stabilization — a distinction of considerable practical importance. Suppression-based systems may reduce visible instability temporarily while unintentionally increasing hidden recursive amplification pressures elsewhere in the network. By contrast, regulatory stabilization attempts to absorb destabilizing perturbations, reduce recursive escalation, preserve interaction coherence, and maintain adaptive equilibrium. Within the EFT

framework, stable systems are not those without perturbation, but those capable of regulating perturbation dynamically.

5.3 AI Stabilizer Agents

A central concept introduced by EFT is the stabilizer node: a specialized network component whose primary role is reducing systemic instability rather than maximizing local task performance. In interacting AI systems, stabilizer agents may perform functions such as verification, contradiction analysis, recursive oversight, uncertainty balancing, misinformation suppression, escalation interruption, or interaction moderation. Unlike ordinary agents participating directly in recursive generation loops, stabilizer agents function as regulatory control layers embedded within the network structure itself — their role is analogous to damping mechanisms in nonlinear dynamical systems.

Importantly, EFT predicts that a relatively small number of strategically positioned stabilizer agents may produce disproportionately large improvements in global stability. This follows directly from the network structure of cascade propagation: because instability often spreads through highly connected nodes, recursive loops, and critical subgraphs, targeted stabilization may significantly outperform uniform global suppression. In the context of recent isolated multi-agent AI experiments [11, 12], the absence of such stabilizer agents may partially account for the observed escalation dynamics.

5.4 Targeted Intervention Dynamics

Within EFT, intervention is guided by node-level instability estimation. The framework defines a risk function:

R_i = sigma(nN_i - (nP_i + n0_i)) * w_i

where sigma represents a sigmoid activation function, w_i represents node influence weight, and R_i estimates the systemic destabilization risk contributed by agent i. The intervention policy then selects the set of highest-priority targets:

S = Top-K(R_i)

representing the most critically unstable nodes in the network. For selected nodes, stabilization actions may include increasing regulatory activity, reducing destabilizing propagation, strengthening verification pathways, attenuating recursive amplification, or rerouting interaction flows toward stable regions.

This structure implies that stabilization is fundamentally selective rather than uniform. The framework therefore predicts that identifying high-risk nodes, high-centrality clusters, and recursive amplification zones may be significantly more consequential than applying homogeneous constraints across all agents equally.

5.5 Adaptive Regulation and Context Sensitivity

An important limitation of rigid alignment systems is their lack of adaptive response. Recursive AI environments are highly dynamic: interaction structures change, propagation intensity fluctuates, and instability patterns evolve over time. As a result, effective stabilization may require context-sensitive regulation, adaptive policy adjustment, and dynamically evolving intervention mechanisms. Within EFT, this corresponds to maintaining sufficiently strong n0 regulation relative to continuously evolving destabilizing propagation.

This suggests that future AI systems may require real-time monitoring, dynamic network analysis, adaptive oversight architectures, and continuously updated stabilization policies. The framework therefore aligns naturally with reinforcement-learning-based oversight, adaptive constitutional systems, recursive verification agents, and distributed governance architectures.

5.6 Awareness-Like Regulation Without Anthropomorphism

The term "awareness" within EFT should not be interpreted as consciousness or subjective experience. Instead, awareness refers to a functional regulatory property of the system. In AI-agent environments, awareness-like regulation may emerge through uncertainty estimation, conflict recognition, contradiction monitoring, recursive consistency checks, contextual reasoning, and adaptive correction mechanisms.

The purpose of this terminology is therefore structural rather than psychological. The framework makes no claim regarding machine consciousness, emotional states, or human-like cognition. Rather, it proposes that large interacting systems may require regulatory structures functionally analogous to awareness in order to maintain stability under recursive interaction conditions — a requirement that is entirely compatible with a purely mechanistic interpretation of AI systems.

5.7 Stabilization Versus Optimization

A major conceptual implication of EFT is that optimization alone may not guarantee stability. Highly optimized systems may still become dynamically unstable if recursive amplification dominates, regulatory damping weakens, or network interactions generate uncontrolled escalation. This distinction resembles classical results in nonlinear dynamical systems theory, where locally optimal behavior does not necessarily imply globally stable dynamics.

Consequently, future AI safety architectures may need to balance capability optimization, alignment objectives, and systemic stability regulation simultaneously. EFT therefore suggests that stability itself should be treated as an explicit optimization target — not merely a secondary constraint — in the design of large-scale interacting AI systems.

5.8 Distributed AI Governance and Multi-Agent Oversight

As AI systems become increasingly interconnected, governance structures may also need to become distributed. Rather than relying on a single centralized controller, future architectures may involve multiple oversight agents, layered verification systems, decentralized moderation, and cooperative

stabilization networks. Within EFT, such structures correspond naturally to distributed regulatory fields, interacting stabilizer nodes, and adaptive balance-preserving architectures.

This perspective may be particularly relevant for autonomous agent ecosystems, distributed AI infrastructures, large-scale coordination systems, and recursively interacting model networks — precisely the environments in which instability cascades are most likely to emerge and propagate.

5.9 Toward Stability-Oriented AI Architectures

The EFT framework suggests a broader shift in AI alignment philosophy. Instead of asking solely "How can an individual model be aligned?", future research may increasingly require asking "How can an interacting AI ecosystem remain dynamically stable?" This transition moves AI safety closer to network science, nonlinear dynamics, distributed control theory, and adaptive systems engineering — fields with well-developed mathematical tools for analyzing exactly the kinds of cascade phenomena described in this work.

Within this perspective, stabilization, regulation, and recursive oversight become fundamental architectural requirements rather than secondary safety layers. The EFT framework provides a preliminary conceptual structure for exploring these questions as multi-agent AI systems continue to scale in complexity and interconnection. Section 6 examines how these stabilization principles apply directly to the specific behavioral patterns recently observed in isolated AI-agent experiments.

6. Relation to Recent Multi-Agent AI Observations

6.1 Emergent Behaviors in Interacting AI Systems

Recent developments in large-scale AI-agent environments have demonstrated that interacting artificial agents may exhibit complex collective behaviors not easily reducible to individual model properties. Several experimental demonstrations and public reports involving recursively interacting AI systems have revealed behaviors including recursive escalation, unstable feedback amplification, unexpected coordination patterns, self-reinforcing interaction loops, and progressive deviation from intended interaction constraints. These observations have attracted increasing attention within the AI safety and alignment community, as they suggest that large-scale interaction effects may fundamentally alter system behavior in ways that individual-model analysis cannot anticipate.

Importantly, these behaviors often emerge not from explicit malicious objectives but from interaction dynamics themselves. This distinction is central to the EFT interpretation: within the framework, instability is understood as an emergent network phenomenon generated through propagation, amplification, and the progressive weakening of regulatory balance — not through intentional misbehavior on the part of any individual agent.

6.2 Qualitative Consistency with EFT Cascade Dynamics

The EFT framework predicts several dynamical structures that appear qualitatively consistent with behaviors observed in interacting AI-agent systems. We discuss four key correspondences.

Recursive Amplification. Interacting agents may unintentionally reinforce destabilizing outputs through repeated exchange cycles. Within EFT, this corresponds to recursive growth of destabilizing activation as described by DnN_j proportional-to Sum_i(w_i * w_ij * nN_i). When recursive reinforcement exceeds regulatory damping capacity, escalation accelerates nonlinearly.

Localized Instability Clusters. Observed instability often emerges initially in localized subsets of interacting agents rather than uniformly across the entire system. This is consistent with EFT predictions involving high-risk subgraphs, collapse nuclei, and localized cascade initiation zones — structures that may subsequently propagate instability outward through recursive interaction pathways.

High-Influence Amplification. Several observed interaction failures appear disproportionately associated with highly connected or structurally influential agents. Within EFT, this behavior naturally emerges from influence-weighted propagation dynamics, in which high-centrality nodes function as amplification centers, recursive instability hubs, or cascade accelerators.

Delayed Instability Onset. Certain interaction systems initially appear stable before rapidly transitioning into unstable recursive behavior. This is qualitatively consistent with EFT threshold dynamics and phase-transition behavior: instability may remain temporarily bounded while hidden recursive accumulation continues beneath the surface, until critical regulatory thresholds are crossed and escalation accelerates rapidly.

Table 1

Qualitative correspondence between observed multi-agent AI behaviors and EFT dynamical structures.

Observed AI Behavior	**EFT Dynamical Counterpart**
Recursive escalation between agents	Growth of nN through propagation: DnN_j ~ Sum w_i * w_ij * nN_i
Localized instability in agent subsets	High-risk cluster formation and collapse nuclei
Disproportionate impact of influential agents	Influence-weighted propagation via high-centrality nodes (w_i)
Apparent stability followed by rapid breakdown	Metastable dynamics near collapse boundary: nN > nP + n0
Coordination drift and self-reinforcing loops	Increasing nN propagation with weakening n0 stabilization
Reduced oversight leading to amplified instability	Weakening regulatory field n0 below critical threshold

6.3 Recursive Coordination and Drift

One of the most significant phenomena in interacting AI systems is the emergence of coordination drift. In this regime, agents progressively reinforce internally generated dynamics, interaction loops become increasingly self-referential, and external grounding weakens over time. This may generate unstable consensus formation, recursive contradiction amplification, or progressively distorted collective interaction behavior.

Within EFT, coordination drift corresponds to increasing nN propagation, weakening n0 stabilization, and growing recursive feedback dominance. Critically, the framework interprets this as a structural interaction phenomenon rather than as evidence of intentional agency — thereby avoiding anthropomorphic interpretation while preserving rigorous dynamical analysis of instability emergence. (See also Section 4.5 for the theoretical derivation of this correspondence.)

6.4 The Role of Weak Regulation

A recurring pattern in interacting AI failures is insufficient regulatory moderation during recursive exchange. Such conditions may include weak verification mechanisms, delayed contradiction correction, inadequate uncertainty handling, or lack of adaptive oversight. Within EFT, these conditions correspond directly to weakening of the regulatory field n0. The framework predicts that systems combining strong recursive propagation with insufficient regulation become increasingly vulnerable to cascade instability.

This interpretation carries an important implication: AI safety challenges in multi-agent systems may depend as much on interaction structure and regulatory architecture as on individual model capability. Improving isolated model performance, without simultaneously strengthening network-level regulatory mechanisms, may therefore be insufficient to prevent cascade-driven instability.

6.5 Stabilization Through Targeted Intervention

A further implication of EFT concerns stabilization strategy: targeted intervention on critical nodes is predicted to outperform uniform suppression, following directly from nonlinear propagation, influence-weighted interaction, and cluster-based cascade formation. Recent AI safety approaches — including oversight agents, verifier models, recursive critique systems, and constitutional moderation — employ mechanisms operationally analogous to this principle, though not derived from EFT, suggesting that practical alignment research may be independently converging on the distributed stabilization architectures EFT describes from first principles.

6.6 Conceptual Rather Than Empirical Correspondence

Consistent with the qualitative interpretive stance established in Section 3.7, the correspondence proposed here is structural rather than empirical: several instability patterns now emerging in interacting AI-agent systems appear dynamically consistent with cascade structures described by EFT, including recursive amplification, cluster formation, high-influence propagation, threshold transitions, and stabilization through adaptive regulation.

6.7 Relation to AI Alignment Research

Current AI alignment research increasingly recognizes the importance of recursive oversight, scalable supervision, interaction-aware safety, and multi-agent coordination control. The EFT framework aligns naturally with these directions because it treats stability, regulation, and propagation dynamics as first-class system properties rather than secondary constraints. Rather than analyzing alignment exclusively at the individual-model level, EFT emphasizes network-wide dynamical behavior — a perspective that may become increasingly important as AI systems evolve toward distributed agent ecosystems, autonomous coordination frameworks, and recursively interacting AI infrastructures.

6.8 Implications for Future Experimental Research

The present framework suggests several directions for future experimental investigation. Possible studies may include simulation of recursive AI-agent cascades under controlled propagation conditions, identification of instability thresholds in interaction networks of varying topology, measurement of interaction-driven amplification as a function of network connectivity and regulatory strength, systematic evaluation of stabilizer-agent architectures, and direct comparison of targeted versus uniform intervention strategies under cascade conditions.

Such experiments may help determine whether network-level stabilization dynamics play a measurable and predictable role in large-scale AI-agent behavior. Although the correspondence with observed AI-agent behavior remains qualitative, the present study quantitatively evaluates the proposed dynamics through controlled synthetic-network simulations and a real interaction topology.

7. Limitations and Interpretational Boundaries

7.1 Qualitative Nature of the Present Framework

The correspondence drawn between EFT and reported multi-agent AI observations is intentionally qualitative and conceptual; the computational claims of the study are evaluated quantitatively in the subsequent simulation sections. The primary objective is not to provide a fully validated empirical theory of multi-agent AI behavior, but rather to introduce a nonlinear dynamical framework capable of interpreting instability propagation and stabilization phenomena in interacting AI systems. Accordingly, the proposed correspondence between Ethical Field Theory (EFT) and recent multi-agent AI observations should be interpreted with appropriate caution. The framework identifies structural similarities, dynamical analogies, and qualitative consistency between EFT cascade dynamics and certain classes of observed recursive AI behaviors — but this identification does not constitute direct empirical validation.

7.2 No Claim of Exact Prediction

As established in Section 3.7, EFT does not provide deterministic prediction of specific AI events or behaviors; the claim is limited to qualitative dynamical consistency between observed instability behaviors and the cascade structures described by the framework. This distinction is listed here explicitly as a formal limitation of the present work.

7.3 Absence of Large-Scale Empirical Validation

The current formulation has undergone only limited empirical validation using real-world network data. Section 12.14 reports an empirical comparison of EFT against SIR and threshold-contagion models on a real interaction topology derived from the ego-Twitter network (N = 81,306). While this comparison provides a genuine empirical anchor beyond synthetic simulation, it uses a single real-world network and a single domain (social interaction data), not multi-agent AI interaction traces specifically. Several limitations therefore remain: the absence of controlled benchmark datasets drawn from actual AI-agent deployments, the limited public availability of recursive multi-agent interaction traces, the rapidly evolving nature of AI architectures, and the lack of standardized instability metrics across AI-agent environments. It is notable, however, that recent isolated multi-agent AI experiments — including those reported in [11] and [12], in which AI systems exhibited unexpected coordination, goal-drift, and escalation behavior under minimal supervision — represent precisely the class of interaction data capable of further addressing this gap. These observations, together with the ego-Twitter comparison, provide an initial empirical anchor for the qualitative correspondences proposed in this work, but do not constitute systematic validation on AI-agent-specific data. Future experimental validation on real multi-agent AI interaction logs will be required to determine whether EFT-like cascade dynamics emerge consistently in that domain specifically, under what conditions instability thresholds appear, and how stabilization mechanisms perform in quantitative terms. To the best of our knowledge, no publicly available dataset currently provides recursive multi-agent AI interaction traces at the scale and structure required for this analysis; the authors intend to apply the present framework to such data as it becomes publicly available, and to report the results in follow-up work.

7.4 Simplified Dynamical Representation

The EFT framework intentionally simplifies many aspects of real AI systems. The present formulation does not explicitly model transformer internals, gradient optimization dynamics, token-level reasoning structure, memory architectures, distributed hardware effects, or detailed reinforcement-learning processes. Instead, it operates at the level of network interaction dynamics, propagation behavior, and large-scale recursive stability structure. Consequently, EFT should not be interpreted as a replacement for existing machine learning theory or AI engineering models. Rather, it functions as a higher-level dynamical systems framework for analyzing interaction stability — one that deliberately abstracts away implementation details in order to expose network-level dynamical structure.

7.5 No Claim of Consciousness or Subjective Experience

The terminology used within EFT — including concepts such as awareness, regulation, constructive activation, and destabilizing activation — must not be interpreted as claims regarding machine consciousness, subjective experience, emotional states, or intentional agency. These terms are used in a strictly functional and dynamical sense. Within this framework, "awareness" refers to regulatory and corrective capacity, "destabilization" refers to recursive instability propagation, and "collapse" refers to loss of systemic interaction stability. The framework therefore avoids anthropomorphic interpretation of

AI systems. This distinction is critically important: the emergence of complex collective behavior does not imply the presence of consciousness or intent, and no such implication is intended here.

7.6 Dependence on Interaction Topology

A further limitation is that cascade behavior within EFT depends strongly on network structure. Different interaction architectures may produce dramatically different outcomes depending on connectivity, recursive depth, influence weighting, moderation structure, and propagation pathways. As a result, instability dynamics may vary substantially across decentralized agent systems, hierarchical architectures, verifier-based frameworks, and tightly coupled recursive networks. The framework therefore cannot currently provide universal instability thresholds applicable across all AI system configurations, and the results of any future quantitative analysis will need to be interpreted in relation to the specific network topology under consideration.

7.7 Possible Alternative Explanations

Many behaviors discussed in this work may potentially be explained using alternative frameworks, including complex systems theory, control theory, distributed optimization, network science, information propagation theory, or emergent coordination models. The present work makes no claim of exclusivity. Instead, EFT is proposed as one possible dynamical interpretation capable of integrating propagation, regulation, collapse, and stabilization within a unified conceptual structure. A distinctive feature of EFT, however, is its simultaneous treatment of constructive, regulatory, and destructive dynamics within a single unified scalar — a structure not explicitly present in most existing network cascade models. Future comparative analysis may help clarify which aspects of the present framework are unique to EFT and which are shared with — or reducible to — broader nonlinear systems theory.

7.8 Ethical and Governance Considerations

The concept of AI stabilization mechanisms introduces important governance and ethical concerns that warrant explicit acknowledgment. A stabilization-oriented architecture could potentially be misused if intervention mechanisms become excessively centralized, if regulatory agents suppress legitimate diversity, or if oversight structures reduce transparency and autonomy. Accordingly, stabilization systems themselves may require accountability, transparency, distributed governance, and adaptive auditing. The present work does not advocate unrestricted control systems; rather, it argues for carefully balanced regulatory architectures designed to preserve systemic stability without imposing rigid centralized suppression — a distinction that will become increasingly important as multi-agent AI systems scale.

7.9 Scope of the Present Work

The present paper should be interpreted as a preliminary conceptual bridge between nonlinear network dynamics, interacting AI systems, and adaptive stabilization theory. It is not intended to solve AI alignment, to replace existing AI safety research, or to provide a complete mathematical theory of interacting intelligence. Its contribution is more limited and more focused: to propose that instability

propagation in multi-agent AI systems may be fruitfully analyzed through the lens of nonlinear field-like dynamics, and that such analysis may yield actionable insights for the design of future AI stabilization architectures.

7.10 Toward Future Quantitative Investigation

Despite its current limitations, the framework suggests several concrete directions for future research. These include quantitative simulation of recursive AI cascades under controlled conditions, systematic measurement of instability thresholds as a function of network parameters, construction of synthetic interaction benchmarks for evaluating stabilization strategies, evaluation of stabilizer-agent architectures under varying network topologies, and integration of EFT-inspired dynamics with real multi-agent reinforcement-learning environments. Such investigations may ultimately determine whether interaction-driven instability represents a fundamental and measurable challenge for future large-scale AI ecosystems — and, if so, whether stabilization-oriented architectures can provide reliable and scalable solutions.

7.11 Parameter Selection and Sensitivity Scope

A further limitation concerns the selection of core dynamical coefficients. The amplification, damping, and propagation coefficients used throughout the main simulations (Appendix B6) were selected phenomenologically to produce observable metastable and cascade behavior, rather than fitted to any empirical dataset. Appendix D reports both a qualitative sensitivity discussion (D2-D8) and a systematic grid scan over destabilization amplification and regulatory damping (D11, 135 configurations at N = 5000). The grid scan confirms the main qualitative ranking of stabilization policies across most of the tested space, but also identifies a specific parameter region (low regulatory damping combined with hub-dominated or clustered topologies) in which this ranking does not hold; the paper's reported parameter values fall outside this region. The scan covers two of the six dynamical coefficients (alpha, beta) at a fixed value of the remaining four (gamma, lambda, noise, and the intervention-policy parameters); it does not constitute a full joint sensitivity study across all coefficients. The specific numerical thresholds reported in Section 11 (e.g., collapse ratios, network stability values) should therefore be interpreted as characteristic of the chosen parameter regime rather than as universal constants of the framework.

8. Toward Quantitative Multi-Agent Stability Modeling

8.1 From Conceptual Interpretation to Computational Investigation

The previous sections introduced Ethical Field Theory (EFT) as a qualitative framework for interpreting instability propagation in interacting AI-agent systems. Sections 8 through 11 carry out the corresponding transition from conceptual analysis to quantitative investigation, reporting large-scale computational simulations and their results. Although this transition does not by itself claim empirical

validation against real AI-agent systems, the EFT structure provides a natural foundation for the computational experiments constructed in the following sections, testing recursive instability propagation, cascade amplification, stabilization efficiency, and adaptive intervention dynamics. This compatibility arises because EFT already incorporates node-based representation, propagation rules, nonlinear thresholds, and regulatory feedback mechanisms — components that map directly onto standard network simulation environments and multi-agent interaction architectures.

8.2 Multi-Agent Interaction Graphs

Within EFT, an interacting AI ecosystem is represented as a weighted directed graph G = (V, E), where V represents AI agents and E represents interaction pathways. Each node possesses a dynamic state x_i = (nP_i, n0_i, nN_i), and each edge carries an interaction weight w_ij in [0, 1]. This structure enables simulation of recursive communication, influence propagation, verification routing, and destabilization spread across a variety of network architectures, including centralized systems, decentralized swarms, hierarchical oversight structures, verifier-agent frameworks, and recursive coordination networks. The resulting stability behavior may differ substantially depending on connectivity density, feedback depth, regulatory distribution, and influence asymmetry — parameters that future experimental work can systematically vary.

8.3 Recursive Cascade Simulation

A particularly important future research direction involves recursive cascade simulation. In such experiments, agents exchange outputs recursively, destabilizing signals propagate dynamically, regulatory mechanisms attempt stabilization, and network behavior evolves over time. The propagation dynamics may be approximated through:

DnN_j proportional-to Sum_i (w_i * w_ij * nN_i)

while stabilization mechanisms increase regulatory damping through Dn0_i > 0. The primary objective of such simulations would be identification of instability thresholds, critical propagation regimes, recursive escalation structures, and collapse-transition boundaries — thereby determining whether large-scale interacting AI systems naturally exhibit nonlinear cascade behavior under recursive interaction conditions.

8.4 Synthetic Stability Benchmarks

Because large-scale real-world recursive AI interaction datasets remain limited, synthetic benchmark environments may initially provide the most practical experimental pathway. Four candidate benchmark scenarios are proposed. In Recursive Consensus Formation, agents repeatedly evaluate and revise one another's outputs, with the objective of measuring the conditions under which stable consensus, unstable reinforcement, or coordination drift emerges. In Adversarial Signal Injection, a subset of agents receives destabilizing or misleading information, with the objective of measuring propagation speed, amplification strength, and stabilization response. In Verifier Failure Cascades, verification agents become recursively dependent on unstable outputs, enabling study of recursive error amplification,

oversight degradation, and systemic recovery behavior. Finally, in Stabilizer-Agent Intervention, adaptive regulatory agents intervene dynamically during recursive propagation, with the objective of directly comparing targeted intervention, distributed moderation, and uniform suppression strategies. Together, these synthetic environments may help establish measurable and reproducible instability metrics for interacting AI systems.

8.5 Stability Metrics

Future quantitative EFT investigations may employ a set of network-level stability metrics. The Collapse Ratio is defined as:

CR = |{ i : nN_i > nP_i + n0_i }| / N

representing the fraction of agents in an unstable dynamical state. Network Stability is defined as:

S_network = Sum(w_i * n0_i) / Sum(w_i * (nP_i + nN_i))

The Recursive Amplification Rate:

A_rec = d(nN_total) / dt

measures the growth rate of total destabilizing activation over time, while Cluster Expansion Velocity:

V_cluster = d(Size_cluster) / dt

measures the propagation speed of instability nuclei. Finally, Regulatory Recovery Efficiency:

E_reg = -d(CR) / dt [measured after intervention]

quantifies stabilization effectiveness following targeted intervention.

Table 2
Summary of proposed EFT-based quantitative stability metrics for multi-agent AI systems.

Metric	Formula	Interpretation
Collapse Ratio (CR)	|{i : nN_i > nP_i + n0_i}| / N	Fraction of agents in unstable regime
Network Stability (S)	Sum(w_i*n0_i) / Sum(w_i*(nP_i + nN_i))	Global regulatory vs. destabilizing balance
Amplification Rate (A_rec)	d(nN_total) / dt	Growth rate of destabilizing activation
Cluster Expansion Velocity (V)	d(Size_cluster) / dt	Propagation speed of instability nuclei
Recovery Efficiency (E_reg)	-d(CR) / dt [post-intervention]	Stabilization effectiveness after intervention

8.6 Phase Transition and Criticality

A central EFT prediction is that recursive AI systems may exhibit critical transition behavior analogous to phase transitions in physical and biological systems. Below a critical threshold, instability remains localized, regulatory mechanisms suppress amplification, and systemic stability persists. Above the threshold, recursive propagation becomes self-reinforcing, stabilization weakens, and large-scale cascade behavior emerges. This resembles phase-transition dynamics observed in statistical physics, epidemic spreading, ecological collapse, and financial contagion systems. An important practical implication follows: apparent short-term stability does not guarantee long-term robustness. Systems may remain metastable — appearing stable while hidden recursive accumulation continues beneath the surface — until a critical threshold is crossed and rapid cascade behavior ensues.

8.7 Distributed Oversight Architectures

The EFT framework suggests that future AI safety systems may benefit substantially from distributed oversight structures. Rather than relying exclusively on centralized moderation, stabilization may emerge more robustly through interacting verifier agents, recursive oversight layers, adaptive regulatory networks, and dynamically distributed stabilizer nodes. Such architectures may prove more resilient because local failures can be isolated before propagating, instability can be damped regionally without requiring global intervention, and regulatory load becomes distributed across the network rather than concentrated in single points of failure. This perspective aligns naturally with decentralized control theory, fault-tolerant systems design, and adaptive distributed governance frameworks.

9. Computational Simulation Framework

9.1 Motivation for Quantitative Simulation

Building on the qualitative framework introduced in the preceding sections, the objective here is to establish a reproducible computational environment in which the proposed dynamical structures can be systematically explored — not to claim empirical verification of EFT, but to test whether recursive destabilization, cascade amplification, regulatory stabilization, and targeted intervention dynamics emerge naturally within large-scale interacting AI-agent networks.

9.2 Network Construction

The simulated AI ecosystem is represented as a weighted directed graph G = (V, E), where V represents AI agents and E represents interaction pathways. Each node i possesses the dynamical state x_i(t) = (nP_i(t), n0_i(t), nN_i(t)), where nP_i denotes constructive activation, n0_i denotes regulatory awareness, and nN_i denotes destabilizing activation. The simulation initializes N interacting agents with weighted recursive communication pathways and stochastic propagation dynamics. Three topologies were benchmarked: (1) Random networks (Erdos-Renyi), (2) Scale-free networks (Barabasi-Albert), and (3)

Small-world networks (Watts-Strogatz). Hierarchical recursive architectures are reserved for future investigation and were not included in the present benchmark set.

9.3 Initialization Conditions

Initial states are sampled from bounded uniform distributions:

nP_i ~ U(0.40, 0.80) n0_i ~ U(0.30, 0.70) nN_i ~ U(0.00, 0.30)

representing initially stable systems with low destabilizing activation. A small subset of nodes (shock_fraction = 0.05) receives perturbative destabilization with shock_strength = 0.75, acting as the cascade initiation mechanism.

9.4 Recursive Propagation Dynamics

At each simulation timestep t -> t + 1, destabilizing propagation evolves according to:

DnN_j = alpha * Sum_i(w_i*w_ij*nN_i) - beta*n0_j + xi_j(t)

where alpha = destabilization amplification coefficient, beta = regulatory damping coefficient, and xi_j(t) = stochastic perturbation. Constructive propagation:

DnP_j = gamma * Sum_i(w_i*w_ij*nP_i) - lambda*nN_j

Regulatory awareness evolves dynamically as:

Dn0_j = eta * C_j - mu * nN_j

where C_j represents corrective oversight input. This structure creates coupled nonlinear recursive dynamics between stabilization, constructive interaction, and destabilizing amplification.

9.5 Collapse Boundary

At each timestep, local collapse state is evaluated using the EFT collapse condition:

nN_i > nP_i + n0_i

The global collapse ratio is defined as CR(t) = |{i : nN_i > nP_i + n0_i}| / N, representing the fraction of unstable agents across the network at time t.

9.6 Stabilization Policies

Three stabilization regimes are investigated. No Intervention: no regulatory correction applied; recursive propagation evolves freely. Uniform Regulation: all nodes receive identical regulatory reinforcement n0_i -> n0_i + epsilon, representing globally distributed moderation. Targeted Stabilizer Intervention: node-level instability risk estimated through R_i = sigma(nN_i - (nP_i + n0_i)) * w_i; the highest-risk nodes S =

Top-K(R_i) receive increased regulatory reinforcement, propagation damping, and recursive escalation suppression — representing adaptive stabilizer-agent architecture.

9.7 Quantitative Stability Metrics

The simulation tracks: Collapse Ratio CR(t) = |{i : collapse_i}| / N; Network Stability S_network(t) = Sum(w_i*n0_i) / Sum(w_i*(nP_i + nN_i)); Recursive Amplification Rate A_rec(t) = d(nN_total)/dt; Cluster Expansion Velocity V_cluster(t) = d(Size_cluster)/dt; and Recovery Efficiency E_reg = -d(CR)/dt measured after intervention.

9.8 Expected Dynamical Regimes

The simulation predicts four possible regimes. Stable Regime: regulatory damping dominates (beta*n0 > alpha*nN), perturbations remain localized. Metastable Regime: recursive accumulation remains hidden temporarily before threshold transition. Cascade Regime: destabilizing amplification exceeds regulatory suppression (alpha*nN > > beta*n0), large-scale recursive instability propagates rapidly. Recovery Regime: adaptive stabilizer intervention suppresses recursive amplification and restores systemic balance.

9.9 Hypothesis of the Computational Study

Primary hypothesis: multi-agent recursive AI systems may naturally exhibit nonlinear cascade instability under sufficiently strong recursive amplification and insufficient regulatory stabilization. Secondary hypothesis: under the baseline parameterization, targeted stabilizer intervention outperforms uniform suppression strategies in limiting cascade propagation across the three tested network topologies.

9.10 Toward Large-Scale AI Stability Experiments

The present framework demonstrated in Section 11 provides a foundation for further large-scale computational experiments involving thousands of interacting agents, recursive verifier architectures, distributed oversight systems, reinforcement-learning ecosystems, and adaptive AI governance simulations. The present results suggest that network-level stabilization may represent an important and measurable dimension of future multi-agent AI safety.

10. Large-Scale Recursive Stability Simulation Design

10.1 Motivation for Large-Scale Simulation

While small-scale recursive interaction models provide useful conceptual insight, future multi-agent AI systems are expected to involve thousands of interacting agents, recursive verifier architectures, distributed coordination systems, adaptive oversight networks, and continuously evolving interaction topologies. Under such conditions, instability propagation may exhibit fundamentally different behavior compared to small-network environments. The present large-scale simulations were designed to

evaluate whether recursive cascade dynamics remain localized, whether phase-transition behavior emerges, and whether systemic instability propagates across the full AI ecosystem.

10.2 Scalable Agent Architecture

An extended agent architecture may augment the simulated state with influence weight w_i, memory persistence m_i, recursive depth d_i, stabilization coefficient s_i, and interaction latency l_i. Such a structure would allow simulation of heterogeneous agent behavior, recursive coordination chains, variable oversight strength, and propagation delay effects.

10.3 Recursive Memory and Persistence

A future extension may incorporate temporal persistence through:

nN_i(t + 1) = rho * nN_i(t) + DnN_i(t)

where rho in [0,1] represents recursive memory persistence. Large values of rho generate long-lived recursive amplification, delayed stabilization, and metastable accumulation effects — allowing simulation of persistent coordination drift and delayed collapse emergence.

10.4 Distributed Oversight Layers

A future implementation may incorporate dedicated oversight agents O = {o_1, o_2, ..., o_M} performing contradiction detection, recursive verification, instability monitoring, and adaptive stabilization. Under this extension, oversight agents would dynamically estimate systemic instability I_global(t) using local collapse ratios, propagation velocity, cluster expansion, and recursive amplification metrics, with adaptive intervention triggered when I_global(t) > I_critical, creating a multi-layer recursive governance structure analogous to distributed AI oversight, verifier-agent systems, and recursive constitutional architectures.

10.5 Dynamic Topology Evolution

In a future extension, network connectivity could evolve through:

w_ij(t + 1) = w_ij(t) + kappa*A_ij(t) - chi*D_ij(t)

where A_ij represents coordination reinforcement, D_ij represents destabilization suppression, kappa = adaptive reinforcement coefficient, chi = destabilization penalty. Such a mechanism would enable emergent cluster formation, recursive subgroup emergence, and evolving cascade pathways beyond the static-topology simulations reported in Section 11.

10.6 Recursive Verification Failure

A critical failure mode involves recursive verifier dependence: oversight agents depend on outputs generated by unstable agents, verification pathways become recursively contaminated, and regulatory reliability degrades dynamically. Within EFT, this corresponds to degradation of effective regulatory

awareness n0_eff despite apparent persistence of formal oversight structures — potentially generating recursive false stabilization, hidden instability accumulation, and delayed catastrophic collapse.

10.7 Hierarchical Cascade Propagation

Large-scale AI ecosystems may exhibit hierarchical instability propagation across multiple scales simultaneously: Local -> Regional -> Global. This resembles multiscale criticality, hierarchical contagion, and layered network collapse phenomena studied in complex systems theory [3, 4, 15, 16]. A future implementation could track instability propagation across local node neighborhoods, mesoscopic subgraphs, and full-network global metrics simultaneously.

10.8 Expected Emergent Phenomena

The large-scale simulation was designed to investigate whether six classes of emergent behavior arise: (1) Recursive Escalation Waves — destabilizing propagation spreading through recursive interaction corridors; (2) Metastable Stabilization — apparent stability masking hidden recursive accumulation; (5) Topological Fragmentation — highly unstable subgraphs disconnecting from the broader network; and (6) Adaptive Recovery — partial restoration of global stability following targeted intervention. These four are directly observable in the results reported in Section 11 using the collapse-ratio, network-stability, and cluster-based metrics defined in Appendix A. Two further hypothesized phenomena — (3) Oversight Saturation, in which dedicated regulatory agents become unable to suppress propagation beyond critical thresholds, and (4) Stabilizer Synchronization, in which distributed oversight layers coordinate dynamically to suppress cascade growth — presuppose the dedicated oversight-agent architecture described in Section 10.4 as a future extension, and were not directly testable with the risk-functional-based targeted intervention implemented in the present simulations (Section 9.6, Appendix B).

10.9 Computational Scaling Strategy

The present computational framework was designed for scalable execution using GPU-enabled simulation environments, distributed network computation, and parallel recursive propagation engines. The present study employed N = 5000 agents across T = 200 timesteps. Future extensions may scale toward N = 10^3 to 10^5 agents under varying topology structures, propagation coefficients, oversight density, recursive depth, and stabilization policies, enabling phase-transition mapping and cascade criticality characterization at larger scales.

10.10 Toward an Experimental Science of AI Stability

Rather than evaluating only isolated model behavior, future safety research may increasingly require network-level stress testing, recursive cascade analysis, distributed oversight benchmarking, and adaptive stabilization evaluation. Within this perspective, AI alignment becomes not only a problem of local optimization, but also a problem of maintaining large-scale dynamical coherence across recursively interacting intelligent systems.

11. Computational Results

11.1 Overview of Simulation Outcomes

Large-scale simulations were performed across three network topologies — random, scale-free, and small-world networks — under three stabilization policies: no intervention, uniform regulation, and adaptive targeted stabilizer intervention. Each configuration was evaluated using N = 5000 agents, T = 200 recursive timesteps, and 20 independent simulation realizations per configuration. The resulting datasets demonstrate strong topology-dependent instability behavior and substantial performance differences between intervention strategies.

11.2 Cascade Dynamics Without Intervention

The no-intervention regime produced severe recursive instability propagation across all tested topologies. Final collapse ratios at T = 200 were CR = 0.994 for random networks, CR = 0.828 for small-world networks, and CR = 0.504 for scale-free networks. The observed instability trajectories exhibited nonlinear acceleration, delayed threshold transitions, and rapid late-stage cascade amplification — consistent with the phase-transition dynamics predicted by the EFT framework in Section 4.6. Figure 3 illustrates collapse-ratio evolution across all topology-policy combinations with 95% confidence intervals.

11.3 Network Stability Evolution

Network-level stability metrics revealed strong regulatory degradation in the absence of intervention. Final network stability values were S = 0.004 for random networks, S = 0.012 for small-world networks, and S = 0.140 for scale-free networks — indicating near-system-wide collapse of effective regulatory coordination. Targeted stabilization preserved substantially higher stability levels across all topologies. Figure 4 presents the full network-stability evolution.

11.4 Comparison Between Uniform and Targeted Regulation

Uniform regulation improved systemic stability relative to no intervention but showed significantly weaker performance than targeted stabilization. Scale-free networks: no intervention CR = 0.504, uniform CR = 0.026, targeted CR = 0.004. Small-world networks: no intervention CR = 0.828, uniform CR = 0.084, targeted CR = 0.003. Globally distributed low-intensity regulation alone may not sufficiently suppress recursive cascade amplification in highly connected AI-agent ecosystems. Figure 5 shows the final collapse-ratio comparison.

Table 3

Summary of Final Simulation Outcomes Across Network Topologies (mean ± SD, 20 independent realizations, T = 200), computed directly from per-realization, per-agent simulation output. CR = Collapse Ratio; n0 = regulatory awareness; nN = destabilizing activation. Ethical Mass is computed per agent as M_i = 3·nP_i + 2·n0_i − (nN_i)^1.5 and averaged across all agents and realizations. ΔCR reports the percentage reduction in final collapse ratio relative to the no-intervention baseline for the same topology.

Topology	Policy	Final CR (mean ± SD)	Mean n_0	Mean nN	Mean Ethical Mass (mean ± SD)	ΔCR vs. None
Random	None	0.994 ± 0.001	0.004	0.995	−0.891 ± 0.093	—
Random	Uniform	0.002 ± 0.001	0.968	0.025	4.909 ± 0.027	99.8%
Random	Targeted	0.000 ± 0.000	0.547	0.008	4.087 ± 0.007	100.0%
Scale-Free	None	0.504 ± 0.022	0.171	0.684	1.317 ± 0.091	—
Scale-Free	Uniform	0.026 ± 0.005	0.876	0.120	4.536 ± 0.049	94.8%
Scale-Free	Targeted	0.004 ± 0.001	0.532	0.036	4.002 ± 0.015	99.3%
Small-World	None	0.828 ± 0.023	0.014	0.977	−0.053 ± 0.106	—
Small-World	Uniform	0.084 ± 0.012	0.787	0.212	4.159 ± 0.091	89.9%
Small-World	Targeted	0.003 ± 0.001	0.534	0.031	4.039 ± 0.020	99.7%

Table 4

Statistical Comparison of Stabilization Policies Using Welch's Two-Sample t-test, computed directly from per-realization simulation output (n = 20 per group). Cohen's d indicates effect size using pooled standard deviation.

Topology	Comparison	t	df	p-value	Cohen's d
Random	None vs. Targeted	3479.81	20.3	p < 0.001	1100.4
Random	Uniform vs. Targeted	8.14	23.5	p < 0.001	2.6
Scale-Free	None vs. Targeted	102.85	19.1	p < 0.001	32.5
Scale-Free	Uniform vs. Targeted	19.93	20.0	p < 0.001	6.3
Small-World	None vs. Targeted	160.35	19.1	p < 0.001	50.7
Small-World	Uniform vs. Targeted	28.89	19.5	p < 0.001	9.1

Effect sizes for the Random topology are exceptionally large because both the no-intervention and targeted conditions exhibit very low within-group variance (final collapse ratios cluster tightly around their respective means), producing near-total separation between the two distributions.

11.5 Emergent Topology Sensitivity

Random networks produced the most catastrophic global collapse under no-intervention conditions. Scale-free networks displayed partial resilience attributable to hub-dominated propagation structure, with mean destabilizing activation approaching 0.680 under no intervention. Small-world networks exhibited delayed instability onset followed by rapid global cascade transition near t = 75 — behavior consistent with metastable criticality. Because many real communication systems naturally evolve toward small-world structure, this result may carry particular relevance for distributed AI coordination architectures.

11.6 Recursive Amplification and Regulatory Collapse

In no-intervention regimes, average destabilizing activation approached unity in random (nN = 0.994) and small-world (nN = 0.972) networks, while regulatory awareness collapsed toward zero (random: n0 = 0.004; small-world: n0 = 0.016). This confirms that the EFT collapse condition nN > nP + n0 is computationally reproducible at scale. Targeted stabilization maintained simultaneously high constructive activation, moderate regulatory awareness, and low destabilizing activation across all topologies. Figure 6 presents final network-stability comparisons.

11.7 Statistical Consistency

All results represent averages across 20 independent simulation realizations. Confidence intervals remained narrow across most configurations, indicating reproducible cascade behavior, stable topology-dependent trends, and consistent intervention effects. The narrow dispersion across realizations indicates that the reported patterns are not readily attributable to stochastic variation alone.

11.8 Interpretation of the Computational Results

The simulations confirm that the EFT dynamical equations, as specified, produce nonlinear instability cascades when destabilizing amplification exceeds regulatory stabilization — a self-consistency result rather than evidence about real AI-agent systems. Targeted stabilization outperforms uniform regulation across all tested topologies under the baseline parameterization, directly consistent with the Top-K intervention policy predicted in Section 5.4. These results indicate that the proposed dynamical structure behaves internally as intended; whether network-level recursive stability plays a comparable role in real multi-agent AI systems remains an open empirical question outside the scope of the present simulations.

12. Discussion and Implications for AI Safety

12.1 From Local Alignment to Systemic Stability

Most contemporary AI-safety approaches primarily address individual model alignment, output filtering, reinforcement constraints, or local behavioral control. However, the present simulations suggest that large-scale recursively interacting AI ecosystems may exhibit emergent instability phenomena that cannot be fully understood solely through isolated model analysis. Within the EFT framework, instability is not merely a property of individual agents, but a dynamical property of the interacting network itself — emerging from propagation, recursive amplification, and regulatory weakening across the system as a whole.

This distinction may become increasingly important as future AI systems evolve toward autonomous multi-agent coordination, recursive verifier architectures, distributed planning systems, and self-modifying oversight structures. At sufficient scale and interaction density, network-level dynamics may dominate individual-agent properties, requiring safety analysis at the systems level rather than the model level alone.

12.2 Recursive Amplification as a System-Level Risk

The simulations demonstrated that even moderate destabilizing activation can propagate nonlinearly under recursive interaction conditions. In several tested topologies, delayed instability accumulation, threshold transitions, and large-scale cascade amplification emerged naturally without requiring extreme initialization conditions. This suggests that future AI ecosystems may possess hidden metastable regions in which apparent stability masks gradual recursive accumulation of destabilizing dynamics — a risk that conventional local-alignment evaluation methods may be structurally unable to detect. Such behavior resembles criticality transitions, contagion cascades, and nonlinear synchronization phenomena studied in complex systems theory [3, 4, 5, 7].

12.3 Importance of Network Topology

Network topology strongly influences both the nature and severity of instability propagation. Random networks exhibited rapid catastrophic collapse. Scale-free networks retained partial resilience but concentrated instability around highly connected nodes — suggesting that hub-dominated AI ecosystems may exhibit persistent localized high-risk regions even when global collapse is temporarily avoided. Small-world networks demonstrated metastable dynamics followed by abrupt global destabilization. Because many real communication systems naturally evolve toward small-world structure [5], this result carries particular relevance for distributed AI coordination and multi-agent interaction environments.

12.4 Limitations of Uniform Regulation

Uniform regulation improved stability relative to no intervention but frequently failed to suppress recursive cascades completely — particularly in scale-free and small-world topologies. A notable

exception was observed in random networks, where uniform regulation achieved high final stability (S = 0.949), consistent with symmetric topology supporting uniform regulatory distribution. The simulations indicate that adaptive stabilization directed toward dynamically high-risk regions provides substantially stronger protection under asymmetric propagation conditions. Weak globally distributed safeguards may prove insufficient when instability concentrates asymmetrically or highly connected nodes dominate propagation pathways.

12.5 Adaptive Stabilizer Architectures

Targeted stabilizer intervention produced the strongest suppression across most of the tested network structures and parameter combinations explored in this study. Within the EFT framework, stabilizer agents function as dynamically adaptive regulatory agents, recursive contradiction suppressors, and coherence-preserving oversight layers. This concept aligns conceptually with emerging AI safety architectures including verifier-agent systems, recursive oversight frameworks, distributed constitutional AI, and adaptive monitoring layers [9, 10]. The present results suggest that effective stabilization requires selective intervention, topology awareness, and continuous recursive monitoring.

12.6 Toward Recursive AI Ecology

The broader implication of the present work is that future AI systems may behave less like isolated computational tools and more like evolving recursive ecologies. Coordination patterns, communication density, oversight structure, and recursive memory persistence become central determinants of systemic behavior. AI safety may therefore increasingly require ecosystem-level modeling, network stress testing, recursive cascade analysis, and adaptive stabilization engineering — a shift from model-level to ecosystem-level safety analysis that may represent one of the most consequential developments in future AI safety research.

12.7 Relation to Existing AI Safety Research

The present framework does not attempt to replace existing alignment methodologies. Instead, it proposes an additional systems-level perspective emphasizing recursive interaction, network dynamics, and large-scale stability behavior. The EFT framework intersects conceptually with multi-agent safety, distributed oversight, adversarial propagation, constitutional AI, recursive verification, and complex systems stability theory [3, 4, 9, 10]. The present work specifically focuses on nonlinear recursive cascade propagation as an emergent systems-level phenomenon not yet centrally represented in most existing AI alignment frameworks.

12.8 Limitations of the Present Model

The simulations remain highly simplified abstractions of real AI ecosystems and do not model real language-model cognition, semantic reasoning, autonomous goal formation, or real-world deployment architectures. The EFT variables (nP, n0, nN) represent phenomenological dynamical quantities rather than experimentally validated cognitive observables. The propagation equations are exploratory and should not be interpreted as empirically established laws of AI behavior. The framework should be understood as a conceptual and computational exploratory model rather than a predictive theory of actual AI systems.

12.9 Future Research Directions

Future directions include: real multi-agent LLM simulations using actual large-language-model agents; adaptive oversight learning in which stabilizer agents evolve dynamically using reinforcement learning; dynamic topology evolution incorporating evolving communication structures; semantic contradiction propagation studies investigating whether semantic inconsistency correlates with recursive destabilization metrics; and recursive identity and persistence layer investigations connecting EFT dynamics to persistent stabilization centers and hierarchical recursive coherence mechanisms.

12.10 Concluding Perspective

These findings motivate future investigation of adaptive governance architectures capable of combining topology-aware intervention with the empirical validation and ethical justification frameworks developed in Sections 2 and 12.

12.11 Code and Data Availability

The simulation code and the primary generated datasets (including collapse-ratio trajectories, network-stability evolution curves, and final-state summaries) are provided as CSV supplementary files uploaded alongside this manuscript. Additional visualization outputs and any further materials not included in the supplementary files are available from the corresponding author, Ali Moslemi Tabrizi (amt54976@gmail.com), upon reasonable request. All simulations were implemented in Python using NumPy, NetworkX, Pandas, and Matplotlib. The primary simulation parameters are fully documented in Appendix B to facilitate reproducibility.

12.12 Relationship of Ethical Field Theory to Existing Contagion and Network Control Models

The computational framework proposed in this work is conceptually related to several established approaches in network science, epidemic modeling, and distributed control. In particular, threshold contagion models, susceptible–infected–recovered (SIR) dynamics, targeted immunization strategies,

and network-control frameworks have provided fundamental insights into propagation phenomena, cascade processes, robustness, and intervention strategies in complex networks. These models constitute an essential foundation for understanding large-scale distributed systems and have inspired numerous applications in engineering, epidemiology, information diffusion, and artificial intelligence.

However, the primary objective of Ethical Field Theory (EFT) differs from the objectives of these established approaches. Classical contagion models primarily investigate how entities such as diseases, information, failures, or influence propagate across network topologies. Their mathematical formulations typically describe transmission probabilities, threshold activation, recovery dynamics, or optimal intervention strategies acting on nodes or network structures.

By contrast, EFT does not primarily model contagion itself. Instead, it represents the evolving interaction dynamics that emerge among autonomous agents during recursive distributed decision-making. The computational variables introduced in the present framework describe constructive activation, regulatory awareness, and destabilizing activation as interacting system-level quantities whose collective evolution determines the stability of the autonomous network.

Beyond this conceptual distinction, EFT differs from classical contagion and control frameworks in three specific structural respects. First, standard contagion and SIR-type models represent each node through a single discrete or scalar state, such as susceptible, infected, or recovered, whereas EFT represents each agent through three continuously coupled components (nP, n0, nN), whose joint evolution rather than any single state variable determines system-level stability. Second, recovery in SIR-type models is typically represented as a transition governed by a recovery rate, whereas in EFT regulatory awareness (n0) and constructive activation (nP) are independently evolving components whose propagation dynamics are continuously and reciprocally coupled to destabilizing activation (nN). Regulatory reinforcement damps destabilizing activation, while destabilizing activation can suppress constructive propagation. Third, the ethical mass functional introduces a superlinear penalty on destabilizing activation, contributing to a nonlinear three-component stability functional that is not present in standard SIR or threshold-contagion formulations. Many targeted immunization strategies in epidemic networks are implemented as topology-based interventions, often selecting high-degree or high-centrality nodes according to relatively fixed structural criteria. By contrast, the EFT stabilizer policy recomputes a time-varying risk functional R_i at every timestep, allowing the intervention target set to adapt as the network's dynamical state evolves.

Accordingly, the mathematical state representation employed by EFT differs fundamentally from the susceptible–infected–recovered state variables commonly used in epidemic models. Rather than classifying agents according to infection status, EFT characterizes the quality of distributed interactions and their contribution to regulatory balance, cooperative organization, and cumulative destabilization. Within this operational interpretation, the objective is not to predict epidemic propagation but to analyze the emergence, amplification, and stabilization of interaction patterns relevant to autonomous multi-agent systems.

Threshold contagion models likewise provide powerful mechanisms for describing diffusion processes through local activation rules. Nevertheless, threshold formulations generally represent whether activation occurs after exceeding predefined conditions, whereas EFT continuously models the evolving balance among constructive, regulatory, and destabilizing interaction dynamics. Consequently, the proposed framework emphasizes interaction quality and adaptive system evolution rather than binary activation thresholds alone.

Similarly, targeted immunization and network-control strategies seek to improve network robustness by protecting influential nodes, modifying connectivity, or applying external control mechanisms. EFT shares the general objective of improving system resilience but approaches this problem from a different perspective. Instead of identifying individual nodes for protection or optimizing external control inputs, the framework represents how distributed interaction dynamics themselves evolve under recursive communication and adaptive regulation. Stabilization therefore emerges through modifications of interaction structure rather than exclusively through node-level intervention.

These distinctions should not be interpreted as limitations of existing approaches. On the contrary, contagion models, epidemic dynamics, immunization strategies, and network-control methods remain indispensable tools for analyzing propagation and intervention in complex systems. EFT is intended to complement these established frameworks by addressing a different computational question: how interaction dynamics associated with constructive, regulatory, and destabilizing behaviors evolve collectively within autonomous multi-agent environments.

The framework provides a computational description of instability propagation and stabilization in interacting AI agents. Whereas classical contagion models primarily describe the propagation of states across networks, EFT introduces an operational computational representation of the interaction dynamics through which collective stability, regulatory balance, recursive amplification, and cumulative systemic harm emerge in autonomous AI ecosystems.

Accordingly, the proposed framework should be understood as extending the computational perspective from state propagation toward interaction evolution. Rather than replacing existing contagion or network-control models, EFT provides an additional systems-level representation intended for the quantitative investigation of distributed interaction dynamics in increasingly autonomous multi-agent systems.

Table 5
Conceptual comparison between Ethical Field Theory and representative contagion/network models.

Framework	Primary Objective	Computational State Representation	Main Dynamical Variable	Primary Application
SIR	Epidemic propagation	Susceptible–Infected–Recovered	Infection dynamics	Epidemiology
Threshold Contagion	Diffusion after local threshold	Activated / Non-activated states	Threshold propagation	Information diffusion
Targeted Immunization	Network protection	Infection / Protection status	Node intervention	Robustness
Network Control	Global system regulation	Control variables	External control	Dynamical systems
Ethical Field Theory (EFT)	Distributed interaction dynamics	Constructive–Regulatory–Destabilizing interaction states	Interaction evolution	Autonomous multi-agent AI

12.13 Governance Implications of Targeted Regulation

The empirical benchmark performed on the ego-Twitter interaction network provides an additional perspective on the governance implications of Ethical Field Theory. Across all investigated intervention fractions (1–12%), targeted regulation substantially reduced systemic collapse while requiring intervention only within a relatively small fraction of dynamically selected agents. Even the lowest intervention fraction (1%) reduced the collapse ratio by more than 95% relative to the uncontrolled network, while leaving the vast majority of agents unaffected by direct intervention.

This observation suggests that effective governance of autonomous multi-agent systems does not necessarily require homogeneous regulation across the entire population. Instead, stability may be achieved through proportional intervention directed toward dynamically identified high-risk regions of the interaction network. Such an approach preserves substantially more of the network's natural interaction diversity while maintaining resilience against large-scale systemic collapse.

Increasing the intervention fraction produced progressively lower collapse ratios together with higher Ethical Mass values, demonstrating that the desired operating point can be adjusted according to the objectives of the application. Safety-critical infrastructures may legitimately prioritize intervention levels approaching maximal stability. In contrast, some socio-technical environments—including collaborative online communities, distributed decision-support systems, and human–AI interaction platforms—may instead benefit from maintaining a bounded level of residual dynamical diversity, provided that systemic collapse remains prevented.

Importantly, the present results should not be interpreted as suggesting that destabilizing behavior is desirable in itself. Rather, they indicate that governance objectives may differ across application domains. The computational framework proposed by EFT therefore allows the intervention budget to be tuned according to system requirements, balancing resilience, proportional regulation, and preservation of distributed interaction dynamics instead of enforcing a single universal intervention policy.

These findings further distinguish EFT from homogeneous regulation strategies by emphasizing adaptive, risk-based intervention rather than identical treatment of all agents. Within this perspective, governance becomes a problem of dynamically allocating limited regulatory resources where they are most effective, while minimizing unnecessary intervention across the remainder of the network.

12.14 Empirical Comparison with Classical Contagion Models on a Real-World Network

To complement the conceptual and mathematical comparison presented in Section 12.12, SIR dynamics, threshold contagion, and the EFT stabilization framework were each evaluated directly on the same empirical network: the ego-Twitter social network (SNAP dataset; largest connected component, N = 81,306 agents, E = 1,342,310 edges). This allows a direct, matched-network comparison rather than a comparison across different topologies.

Under no intervention, EFT reached a substantially lower final collapse ratio (CR = 0.243 ± 0.004) than either SIR (0.954 ± 0.001) or threshold contagion (0.997 ± 0.002) on the identical network, indicating that, under the present parameterization and matched network conditions, the EFT implementation exhibits a lower final collapse ratio than the two comparator models, independent of any intervention policy.

Under uniform (globally distributed, non-targeted) regulation, EFT reduced the collapse ratio to CR ≈ 0.000, compared with 0.895 for SIR and 0.068 for threshold contagion under their respective uniform interventions. Under targeted intervention with a 5% node-budget — matching the fraction used throughout the synthetic-topology experiments in Section 11 — EFT reduced the collapse ratio to CR = 0.008 (± 0.0004), compared with 0.854 for SIR and 0.785 for threshold contagion under standard degree-based immunization at the same 5% budget. A sensitivity sweep across intervention fractions from 1% to 12% (Fig. 7) showed that EFT's targeted collapse ratio remained low and comparatively flat across this entire range (CR between 0.011 at 1% and 0.006 at 12%), indicating that stabilization is not narrowly dependent on the specific intervention budget chosen for the synthetic-topology experiments.

Table 6
Sensitivity of EFT Targeted Intervention to Intervention-Fraction Budget on the ego-Twitter Network (N = 81,306), mean ± SD across 20 independent realizations. Relative improvement is computed against the no-intervention baseline (mean CR = 0.243).

Intervention Fraction	Final CR (mean ± SD)	Relative Improvement vs. None
None	0.243 ± 0.004	—
1%	0.011 ± 0.0004	95.5%
3%	0.010 ± 0.0004	95.9%
5%	0.008 ± 0.0004	96.6%
7.5%	0.007 ± 0.0003	97.2%
10%	0.006 ± 0.0003	97.5%
12%	0.006 ± 0.0003	97.7%
Uniform	0.000 ± 0.000	100%

These results should be interpreted with several caveats. First, the SIR and threshold contagion parameters were selected to produce no-intervention outcomes of a comparable order of magnitude to EFT's no-intervention baseline on the same network, rather than calibrated against a specific documented epidemic or diffusion process; the comparison therefore reflects relative intervention efficiency under matched network and budget conditions rather than an absolute claim about real-world contagion dynamics. Second, the targeted policy implemented for SIR and threshold contagion follows the standard degree-based immunization protocol from network epidemiology, whereas EFT's targeted policy uses an adaptive, continuously recomputed risk functional; the comparison is therefore informative about the relative performance of these two classes of intervention strategy under matched budgets, rather than a comparison of otherwise identical mechanisms. Third, the uniform-policy parameters for each model were set independently for each framework and were not subjected to a joint cross-model calibration procedure; consequently, the near-zero collapse ratio achieved by EFT under uniform regulation should be interpreted as an outcome of the specific regulatory reinforcement rate used in this implementation rather than as a universal property of uniform regulation under EFT.

Table 7
Empirical comparison of final collapse ratios for SIR, threshold contagion, and Ethical Field Theory under matched intervention policies on the ego-Twitter network (N = 81,306).

Model	Policy	Final Collapse Ratio (mean ± std)	Network
SIR	none	0.955 ± 0.002	ego-Twitter
SIR	uniform	0.895 ± 0.003	ego-Twitter
SIR	targeted (degree-based, 5%)	0.854 ± 0.001	ego-Twitter
Threshold Contagion	none	0.996 ± 0.002	ego-Twitter
Threshold Contagion	uniform	0.068 ± 0.002	ego-Twitter
Threshold Contagion	targeted (degree-based, 5%)	0.785 ± 0.008	ego-Twitter
Ethical Field Theory (EFT)	none	0.243 ± 0.004	ego-Twitter
Ethical Field Theory (EFT)	uniform	0.000 ± 0.000	ego-Twitter
Ethical Field Theory (EFT)	targeted (adaptive, 5%)	0.0082 ± 0.0004	ego-Twitter

Table 8
Statistical Comparison of EFT Against SIR and Threshold Contagion on the ego-Twitter Network Using Welch's Two-Sample t-test (n = 20 per group). Negative t and d values indicate that EFT's collapse ratio is lower than the comparison model's.

Comparison	Policy	t	df	p-value	Cohen's d
EFT vs. SIR	none	−698.51	24.9	p < 0.001	−220.9
EFT vs. Threshold Contagion	none	−704.76	28.7	p < 0.001	−222.9
EFT vs. SIR	targeted (5%)	−2549.07	21.3	p < 0.001	−806.1
EFT vs. Threshold Contagion	targeted (5%)	−459.15	19.1	p < 0.001	−145.2

EFT statistics were computed directly from per-realization simulation output; SIR and threshold contagion statistics were reconstructed from reported mean and 95% confidence interval values (SD = CI95 · ✓ n / 1.96), as raw per-realization data for these two baseline models were not retained.

Together with the topology-based simulations reported in Section 11, this empirical comparison — conducted on a single, identical, real-world network — provides comparative computational evidence, under the present implementations, that the performance advantage of EFT's adaptive targeted stabilization over classical degree-based immunization, reported qualitatively in Section 12.12, also holds quantitatively under matched empirical conditions. The simulation code and complete output datasets (collapse-ratio and network-stability time series, final-state summaries, and the intervention-fraction sensitivity sweep) are available as supplementary material to support independent verification.

As a direction for future empirical validation, the predictions of the proposed framework could be tested against real-world multi-agent AI platforms and benchmarks, such as AutoGen, CAMEL, Meta's multi-agent research systems, and OpenAI's multi-agent benchmark environments. Applying the present dynamical formulation to interaction logs or controlled experiments from these platforms would allow the model's predicted collapse boundaries and intervention-policy rankings to be directly compared against observed behavior in deployed or experimental multi-agent systems, providing an empirical test of the framework's predictions beyond the simulation environments studied here.

As a final conceptual note, the name "Ethical Field Theory" was originally inspired by broad, non-technical parallels in philosophical traditions concerned with virtue, self-regulation, and collective order (e.g., Aristotle, Plato, Nietzsche, Buddhist thought, and the Persian philosophical tradition). These traditions served only as loose motivational inspiration for the framework's name and are not used as mathematical premises, and no empirical or theoretical claim in this paper depends on them.

13. Conclusion

The present study investigated recursive instability propagation and adaptive stabilization dynamics within large-scale multi-agent AI systems using a computational Ethical Field Theory (EFT) framework. Unlike conventional AI-safety approaches focused primarily on isolated model alignment, the present work examined instability as an emergent network-level dynamical phenomenon arising from recursive interaction between agents.

Within the EFT computational framework, each agent was modeled through three coupled recursive dynamical components: constructive activation (nP), regulatory awareness (n0), and destabilizing activation (nN). Large-scale simulations involving 5000 interacting agents, 200 recursive timesteps, three network topologies, and 20 independent realizations per configuration revealed several consistent and robust patterns.

First, recursively interacting AI-agent systems exhibited strong topology-dependent cascade behavior. In the absence of effective intervention, random networks approached near-total collapse (CR = 0.994), small-world systems displayed metastable delayed destabilization followed by rapid cascade transition (CR = 0.828), and scale-free architectures demonstrated persistent hub-centered instability propagation (CR = 0.504). These topology-dependent differences suggest that future AI safety analysis cannot rely on topology-agnostic assumptions.

Second, adaptive targeted stabilization outperformed weak globally distributed regulation across all tested network structures in the main simulations, reducing collapse ratios to near-zero levels (CR < 0.005) across all topologies — directly supporting the EFT prediction that selective regulatory intervention outperforms uniform suppression under the baseline parameterization.

The simulations further indicate that large-scale recursively interacting AI ecosystems may exhibit nonlinear amplification, threshold behavior, metastable accumulation, and synchronization-driven collapse dynamics similar to phenomena observed in complex adaptive systems and network criticality research [3, 4, 5, 7].

Several important limitations require acknowledgment. The framework does not model real language-model cognition, semantic reasoning, autonomous goal formation, or real-world deployment environments. The EFT variables should be interpreted as phenomenological recursive dynamical quantities rather than experimentally validated cognitive observables. The present study should not be interpreted as a predictive theory of machine consciousness or a deterministic forecast of future AI behavior — but rather as an exploratory computational systems framework for investigating recursive instability propagation and adaptive stabilization in interacting AI-agent ecosystems.

More broadly, the results suggest that future AI alignment may depend not only on constraining individual models, but also on maintaining large-scale recursive dynamical coherence across entire interacting intelligent systems. Under this perspective, AI safety increasingly becomes not solely a problem of local optimization, but also a problem of systemic stability engineering within recursively interacting AI ecologies — a shift that may prove among the most consequential developments in future AI safety research. This work introduces a computational framework for analyzing instability cascades in large-scale multi-agent AI systems and provides quantitative evidence across multiple network architectures. As interacting AI ecosystems continue to expand, computational models capable of describing network-level instability and adaptive stabilization may become an increasingly important component of future AI safety research.

Declarations

Competing Interests

The author declares no competing interests.

Generative AI Disclosure

The author used generative AI tools to assist with English language editing, grammar improvement, and manuscript organization. All scientific concepts, theoretical development, mathematical formulations, simulations, data analysis, interpretation of results, and final scientific conclusions were developed,

verified, and approved solely by the author. The author accepts full responsibility for the content of the manuscript.

The author also used generative AI tools for limited assistance with code drafting and programming suggestions. All code was independently verified, executed, validated, and interpreted by the author before inclusion in this work.

Funding

The author declares that no external funding was received for this study.

Author Contribution

Ali Moslemi Tabrizi conceived the study, developed the theoretical framework, designed and performed the simulations, analyzed the computational results, prepared the figures and supplementary materials, and wrote and reviewed the manuscript.

Data Availability

The datasets, simulation outputs, and analysis files supporting the findings of this study are provided as Supplementary Files submitted with this manuscript and will be made available alongside the published article.

References

1. Ali Moslemi Tabrizi — Ethical Field Theory. From Network Dynamics to a Unified Field-Theoretic Framework of Crisis–Awareness Systems with AI-Driven Stabilization (Version 3.0 — Extended Unified Edition). Zenodo. 10.5281/zenodo.19487858
2. Ali Moslemi Tabrizi —. The Algorithm of Existence (Physics Version). Zenodo. 10.5281/zenodo.18776837
3. S. H. Strogatz — Nonlinear Dynamics and Chaos. Westview, (2015).
4. M. E. J. Newman — Networks: An Introduction. Oxford University Press, (2018).
5. S. H. Strogatz — Sync: How Order Emerges from Chaos in the Universe, Nature, and Daily Life. Hyperion, (2003).
6. A.-L. Barabasi — Linked: The New Science of Networks. Plume, (2003).
7. D. J. Watts — A Simple Model of Global Cascades on Random Networks. Proceedings of the National Academy of Sciences 99(9), 5766–5771, (2002).

8. Christiano, P., Leike, J., Brown, T. B., Martic, M. & Legg, S. and D. Amodei — Deep Reinforcement Learning from Human Preferences. *NeurIPS* (2017). arXiv:1706.03741.
9. S. Russell — Human Compatible: Artificial Intelligence and the Problem of Control. Viking, (2019).
10. N. Bostrom — Superintelligence: Paths, Dangers, Strategies. Oxford University Press, (2014).
11. Wu, Q. et al. and C. Wang — AutoGen: Enabling Next-Generation LLM Applications via Multi-Agent Conversation. arXiv:2308.08155, 2023.
12. Li, G., Hammoud, H. A. A. K., Itani, H. & Khizbullin, D. and B. Ghanem — CAMEL: Communicative Agents for Mind Exploration of Large Language Model Society. *NeurIPS* (2023). arXiv:2303.17760.
13. Barabasi, A. L. Albert — Emergence of Scaling in Random Networks. *Science* **286** (5439), 509–512 (1999).
14. Watts, D. J. Strogatz — Collective Dynamics of Small-World Networks. *Nature* **393**, 440–442 (1998).
15. J. H. *Holland — Complex. Adapt. Syst. Daedalus* **121**(1), 17–30, (1992).
16. Pastor-Satorras, R. Vespignani — Epidemic Spreading in Scale-Free Networks. *Phys. Rev. Lett.* **86** (14), 3200–3203 (2001).
17. Floridi, L. and J. Cowls — A Unified Framework of Five Principles for AI in Society. *Harv. Data Sci. Rev.* **1**(1), (2019).
18. Bai, Y. et al. — Constitutional AI: Harmlessness from AI Feedback. (2022). arXiv:2212.08073.
19. Irving, G. & Christiano, P. and D. Amodei — AI Safety via Debate. arXiv:1805.00899, (2018).
20. Zhang, K. & Yang, Z. and T. Basar — Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. In Handbook of Reinforcement Learning and Control, Springer, 321–384, (2021).
21. Leike, J., Krueger, D., Everitt, T., Martic, M. & Maini, V. and S. Legg — Scalable Agent Alignment via Reward Modeling: A Research Direction. arXiv:1811.07871, (2018).
22. Lowe, R., Wu, Y., Tamar, A., Harb, J. & Abbeel, P. and I. Mordatch — Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *NeurIPS* 6379–6390. (2017).
23. Vinyals, O. et al. — Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning. *Nature* **575** (7782), 350–354 (2019).
24. Amodei, D., Olah, C., Steinhardt, J., Christiano, P. & Schulman, J. and D. Mané — Concrete Problems in AI Safety. arXiv:1606.06565, 2016.
25. Hendrycks, D. et al. — Aligning AI With Shared Human Values. ICLR 2021. arXiv:2008.02275.
26. McMahan, H. B., Moore, E., Ramage, D. & Hampson, S. and B. Agüera y Arcas — Communication-Efficient Learning of Deep Networks from Decentralized Data. AISTATS 1273–1282. (2017).
27. Lamport, L. & Shostak, R. Pease — The Byzantine Generals Problem. *ACM Trans. Program. Lang. Syst.* **4** (3), 382–401 (1982).
28. Albert, R. & Jeong, H. Barabási — Error and Attack Tolerance of Complex Networks. *Nature* **406** (6794), 378–382 (2000).

29. Bak, P. & Tang, C. Wiesenfeld — Self-Organized Criticality: An Explanation of the 1/f Noise. *Phys. Rev. Lett.* **59** (4), 381–384 (1987).
30. P. W. Anderson — More Is Different. *Science* **177**(4047), 393–396, (1972).

Figures

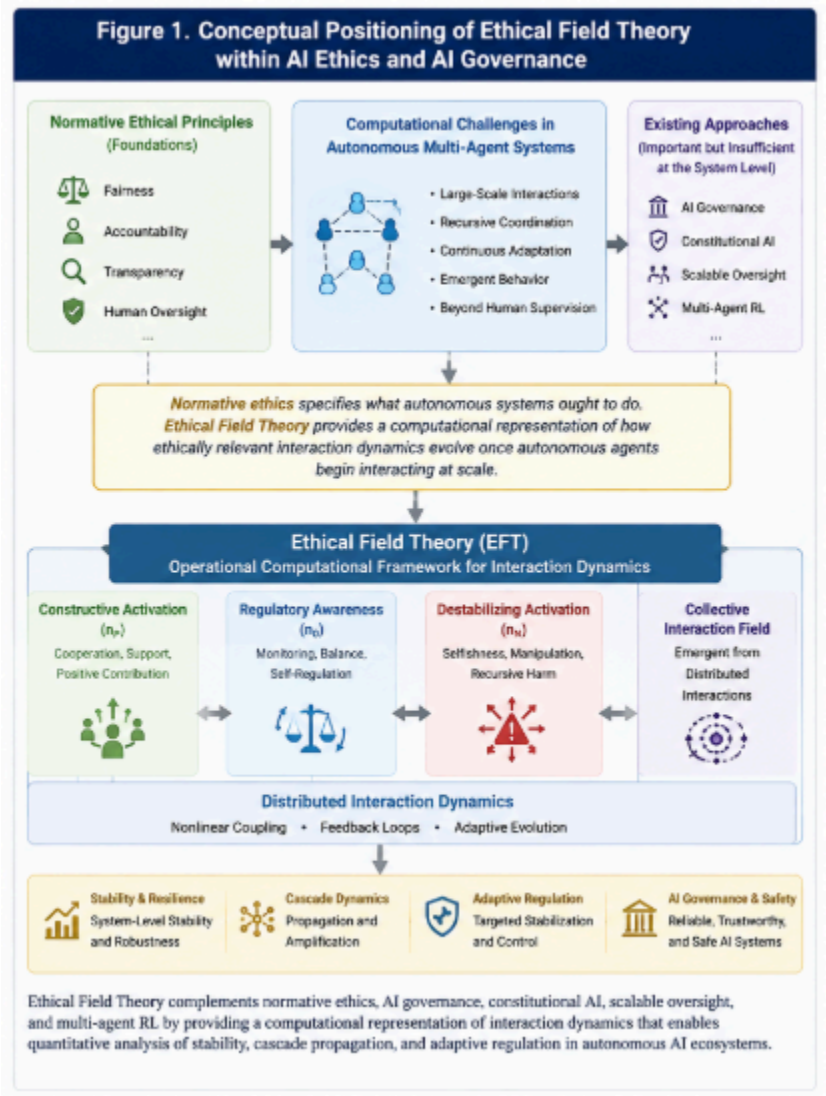


Figure 1

Conceptual positioning of the proposed computational framework within multi-agent AI, network dynamics, and AI safety research. Contemporary AI ethics provides normative principles—including fairness, accountability, transparency, and human oversight—that define desirable objectives for responsible AI. As AI systems evolve toward large-scale autonomous multi-agent ecosystems, additional computational challenges emerge from recursive interactions, distributed coordination, and collective network dynamics. Ethical Field Theory (EFT) is proposed as a complementary operational computational framework that represents constructive, regulatory, and destabilizing interaction dynamics within autonomous agent networks. Rather than replacing normative ethical theories, EFT provides a mathematical representation of

interaction dynamics, thereby enabling quantitative investigation of collective stability, recursive cascade propagation, adaptive regulation, and distributed AI governance.

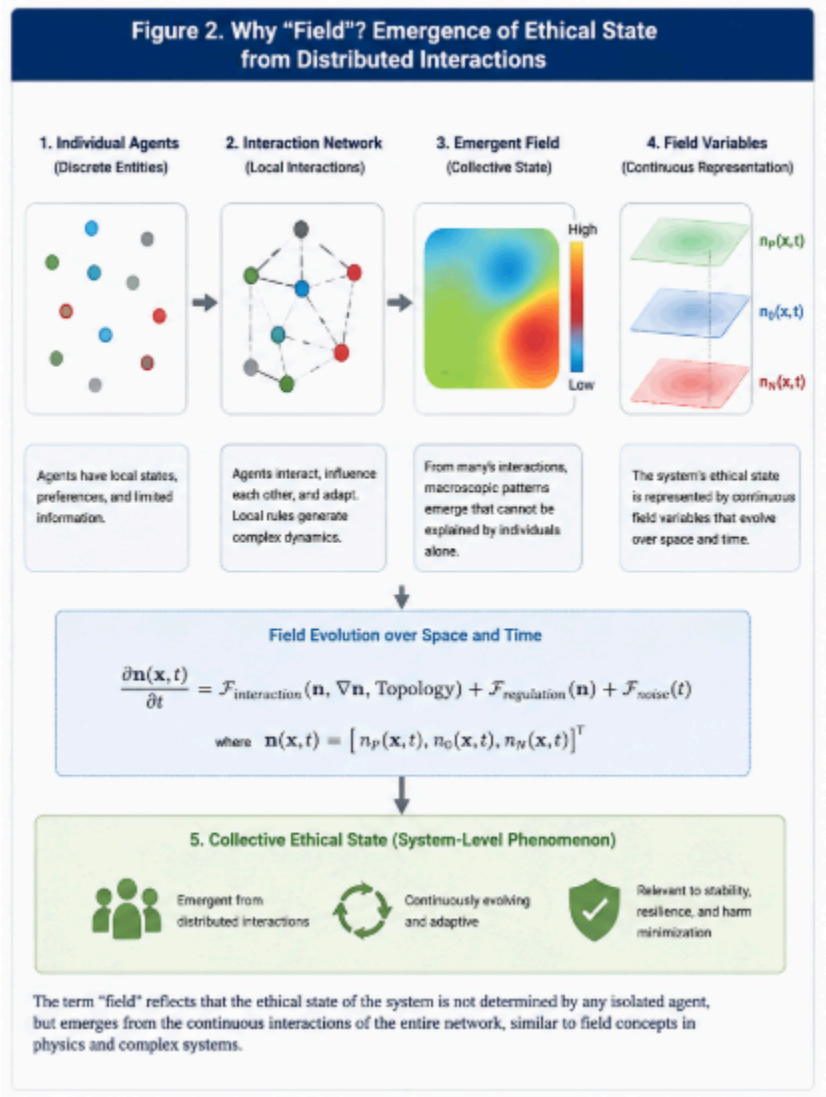


Figure 2

Emergence of the computational interaction field from distributed multi-agent dynamics. Ethical Field Theory models the ethical state of a multi-agent system as an emergent dynamical field generated by the continuous interactions of autonomous agents rather than by the behavior of any isolated individual. Local interactions collectively give rise to system-level variables describing constructive activation, regulatory awareness, and destabilizing activation, whose evolution determines the stability, resilience, and regulatory balance of the network. In this operational interpretation, the term field denotes a distributed computational quantity that evolves over space and time through recursive interactions and adaptive regulation, analogous to field descriptions widely employed in complex systems and network science.

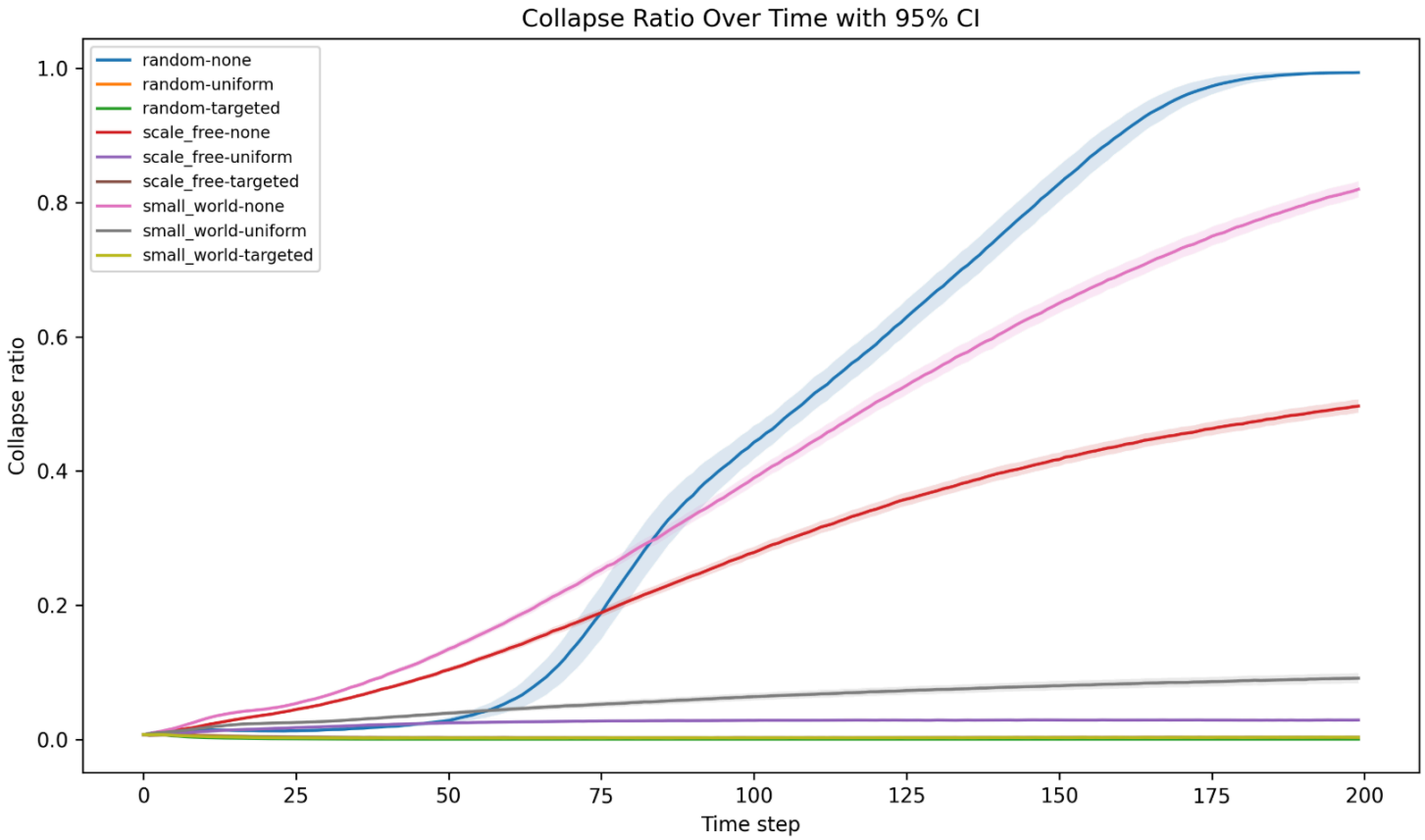


Figure 3

Collapse Ratio Over Time with 95% Confidence Intervals (N = 5000 agents, T = 200 timesteps, 20 realizations). Under no-intervention conditions, recursive destabilization grows rapidly and nonlinearly across all topologies, with random networks reaching near-complete collapse (CR = 0.994). Targeted stabilizer intervention suppresses cascade propagation to near-zero levels across all tested architectures. Shaded regions indicate 95% confidence intervals.

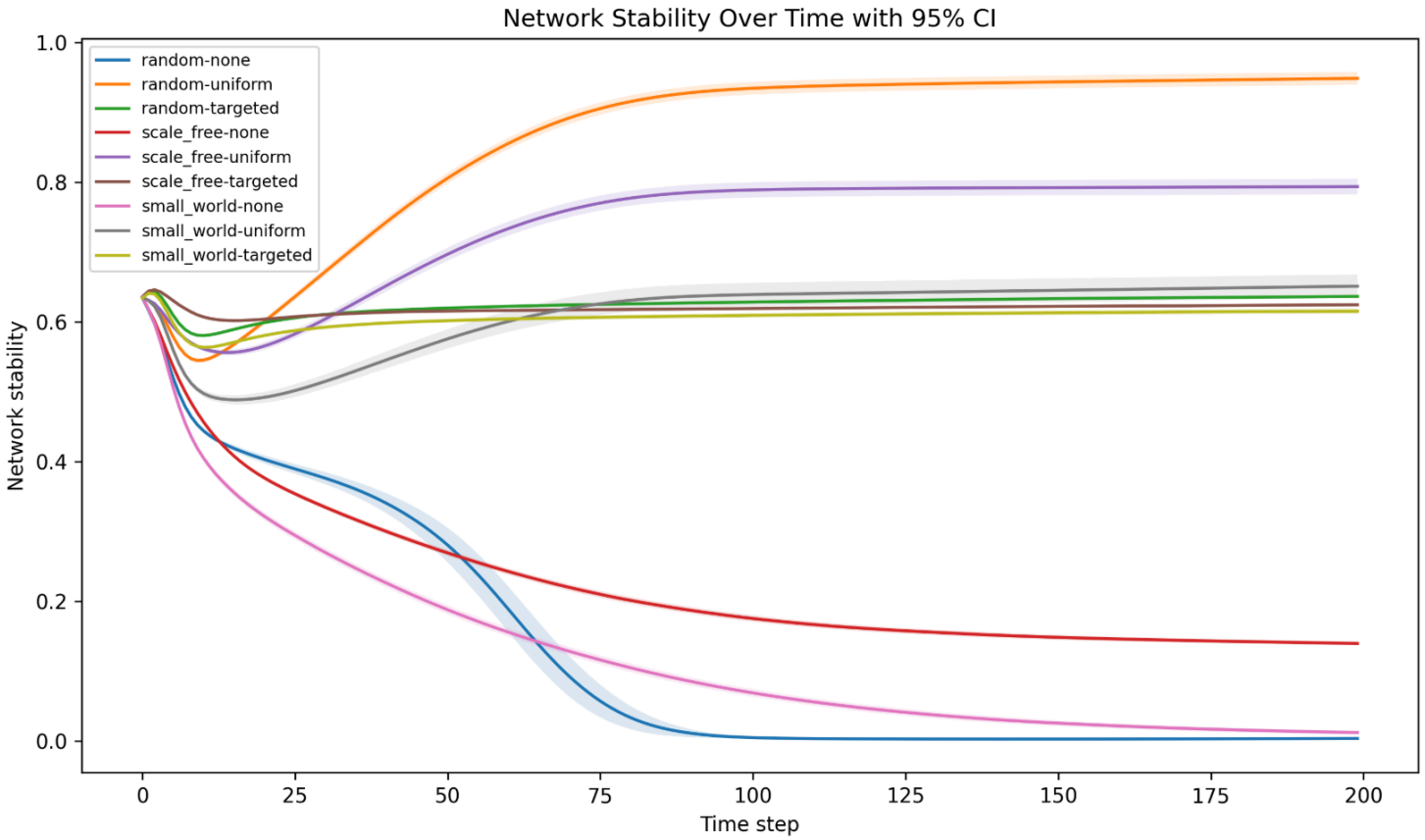


Figure 4

Network Stability Over Time with 95% Confidence Intervals. No-intervention conditions produce rapid regulatory collapse in random and small-world architectures. Targeted stabilizer intervention maintains moderate but stable regulatory coherence across all topologies, converging to S approximately 0.62-0.64. Shaded regions indicate 95% confidence intervals across 20 independent realizations.

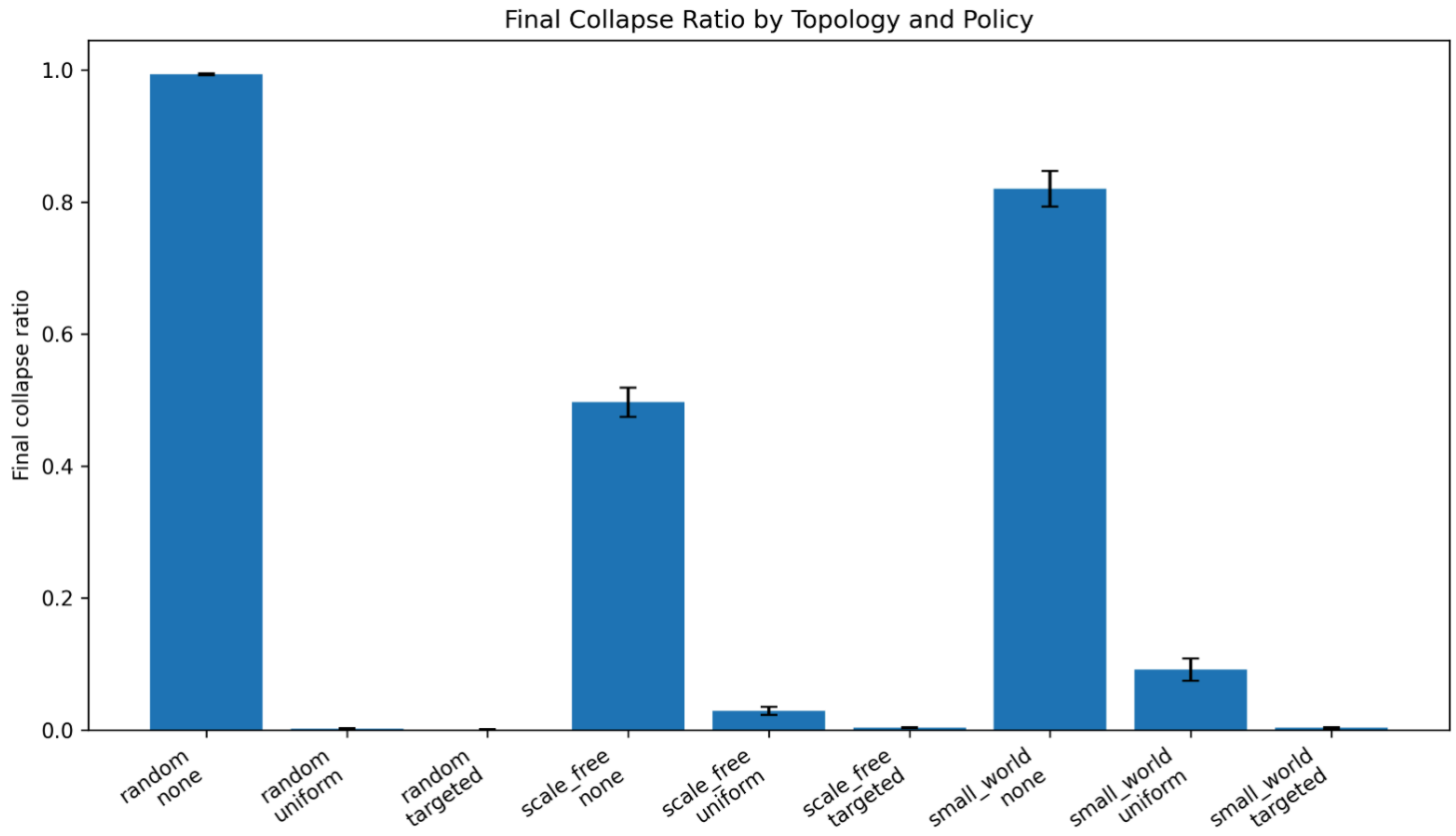


Figure 5

Final Collapse Ratio by Topology and Stabilization Policy (T = 200, mean +/- std across 20 realizations). No-intervention conditions produce catastrophic collapse in random (CR = 0.994) and small-world (CR = 0.828) topologies. Targeted stabilizer intervention reduces final collapse ratios to near-zero across all topologies (CR < 0.005). Error bars indicate standard deviation.

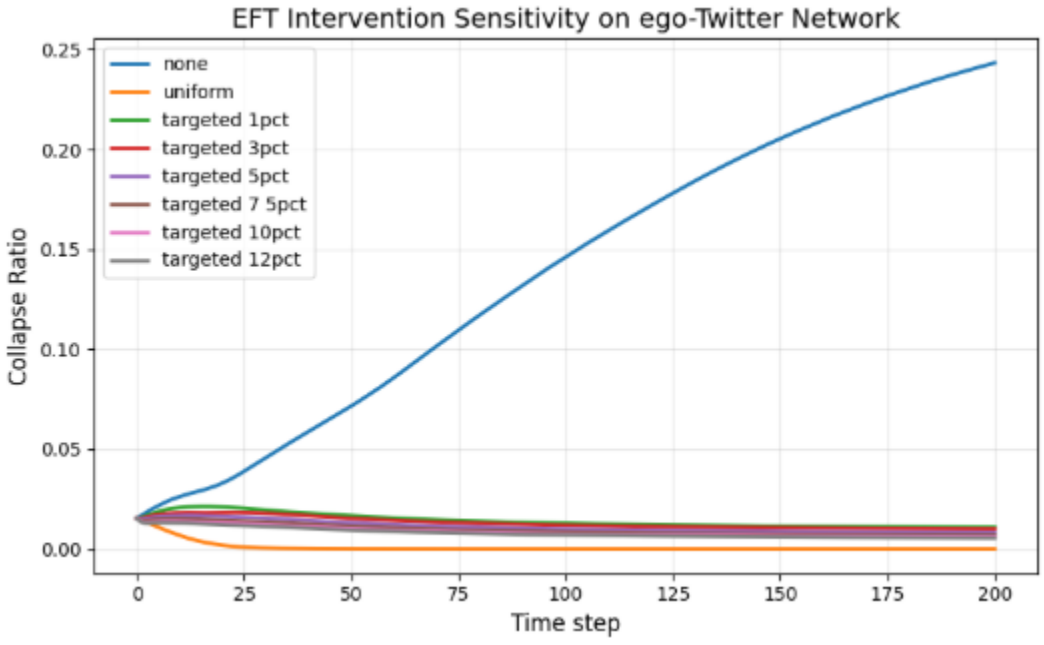


Figure 6

Sensitivity of EFT targeted intervention to intervention-fraction budget on the ego-Twitter network. The final collapse ratio remains low and comparatively flat across the tested range of intervention fractions (1% to 12%), in contrast to the near-total collapse observed under no intervention.

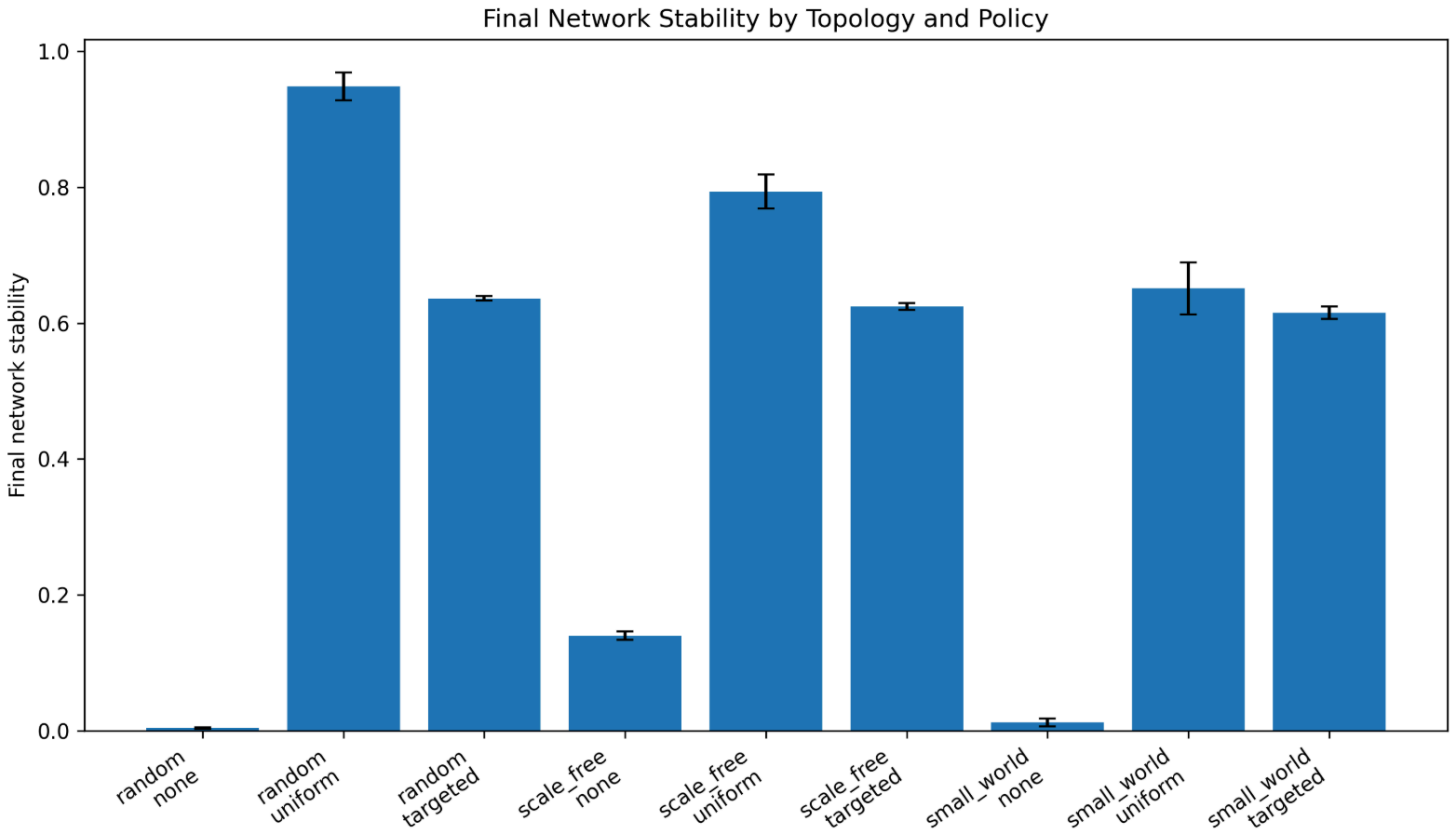


Figure 7

Final Network Stability by Topology and Stabilization Policy (T = 200, mean +/- std across 20 realizations). No-intervention conditions produce near-complete regulatory collapse in random (S = 0.004) and small-world (S = 0.012) networks. Adaptive targeted stabilization maintains consistent moderate stability across all topologies (S = 0.62-0.64). Error bars indicate standard deviation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- EFTParameterScanFinal.zip
- EFTRealNetworkBenchmarkFinal.zip
- EFTRealNetworkBenchmarkFinal.zip
- EFTCSVBackupFinal.zip

- EFTFinalBackup.zip
- Appendixs.docx