

Multi-Agent Systems for Systemic Risk and Resilience

NGO TRUNG KIEN

September 13, 2026

Contents

I	Foundations	57
1	Introduction	59
1.1	Chapter Question	61
1.2	Motivation	61
1.3	Individual Safety vs System Resilience	61
1.4	The Four-Level Framework	63
1.5	What This Book Does and Does Not Claim	64
1.5.1	What this book establishes:	64
1.5.2	What this book does NOT claim:	64
1.5.3	Epistemic status of claims:	65
1.6	Book Structure and Reading Paths	65
1.6.1	Reading paths:	67
1.7	Definitions	67
1.8	Theoretical Mechanism	67
1.9	Evidence	68
1.10	Example	68
1.11	Alternative Explanation	70
1.12	Counterargument	71
1.13	Practical Implication	71
1.14	Limitation	72
1.15	Chapter Summary	72
1.16	Research Questions and Exercises	73
2	Agents and Complex Adaptive Systems	75
2.1	Chapter Question	76
2.2	Motivation	76
2.3	Definitions	77
2.4	Multi-Agent System Architectures	78
2.4.1	Hierarchical Architectures	78

2.4.2	Peer-to-Peer Architectures	78
2.4.3	Blackboard Architectures	78
2.4.4	Market-Based Architectures	79
2.4.5	Hybrid Architectures	79
2.5	Autonomy Spectra and Delegation	79
2.5.1	Autonomy and Speed	79
2.5.2	Autonomy and Scope	80
2.5.3	Delegation Chains	80
2.6	Agent Populations vs Single Agents	80
2.6.1	Emergence	80
2.6.2	Irreducibility	81
2.6.3	The Composition Fallacy	81
2.7	Why Multi-Agent Properties Matter	82
2.8	Example	82
2.9	Alternative Explanation	83
2.10	Counterargument	83
2.11	Practical Implication	84
2.12	Limitation	84
2.13	Chapter Summary	85
2.14	Research Questions and Exercises	85
2.15	Guiding Question for This Section	86
2.16	Motivation	86
2.17	Definitions	86
2.18	Complex Adaptive Systems	87
2.18.1	Properties of CAS	87
2.18.2	CAS Versus Complicated Systems	88
2.19	Emergence: Weak and Strong	88
2.19.1	Weak Emergence	88
2.19.2	Strong Emergence	89
2.19.3	Implications for AI Safety	89
2.20	Non-Linearity and Phase Transitions	89
2.20.1	Non-Linear Dynamics	89
2.20.2	Phase Transitions in Agent Populations	90
2.20.3	Early Warning Signals	90
2.21	Feedback Loops and Amplification	91
2.21.1	Positive Feedback (Destabilizing)	91
2.21.2	Negative Feedback (Stabilizing)	91
2.21.3	Feedback Balance	91
2.22	Self-Organized Criticality	92
2.22.1	The Sandpile Model	92

2.22.2	SOC in Agent Populations	92
2.22.3	Evidence and Limitations	92
2.23	Implications for AI-Agent Populations	93
2.24	Example	94
2.25	Alternative Explanation	94
2.26	Counterargument	95
2.27	Practical Implication	95
2.28	Limitation	96
2.29	Chapter Summary	96
2.30	Research Questions and Exercises	96
3	Network Science and Propagation	99
3.1	Chapter Question	100
3.2	Motivation	101
3.3	Definitions	101
3.4	Network Topologies	102
3.4.1	Random Networks (Erdos-Renyi)	102
3.4.2	Scale-Free Networks (Barabasi-Albert)	103
3.4.3	Small-World Networks (Watts-Strogatz)	103
3.4.4	Modular Networks	104
3.4.5	Topology Comparison	104
3.5	Centrality Measures	105
3.5.1	Degree Centrality	105
3.5.2	Betweenness Centrality	105
3.5.3	Closeness Centrality	105
3.5.4	Eigenvector Centrality	105
3.5.5	PageRank Centrality	106
3.6	Network Robustness and Fragility	106
3.6.1	Node Removal Analysis	106
3.6.2	Percolation Threshold	106
3.6.3	Cascade Models on Networks	106
3.7	Multiplex and Temporal Networks	107
3.7.1	Multiplex Networks	107
3.7.2	Temporal Networks	107
3.8	Why Topology Matters for Propagation	108
3.9	Example	108
3.10	Alternative Explanation	109
3.11	Counterargument	109
3.12	Practical Implication	109
3.13	Limitation	110

3.14	Chapter Summary	110
3.15	Research Questions and Exercises	111
3.16	Guiding Question for This Section	111
3.17	Motivation	111
3.18	Interaction Graph	112
3.18.1	Definition	112
3.18.2	Properties	112
3.18.3	Systemic Risk Relevance	112
3.19	Dependency Graph	112
3.19.1	Definition	112
3.19.2	Properties	113
3.19.3	Systemic Risk Relevance	113
3.20	Information-Flow Graph	113
3.20.1	Definition	113
3.20.2	Properties	113
3.20.3	Systemic Risk Relevance	113
3.21	Authority Graph	114
3.21.1	Definition	114
3.21.2	Properties	114
3.21.3	Systemic Risk Relevance	114
3.22	Multiplex Network Representation	114
3.22.1	Definition	114
3.22.2	Cross-Layer Propagation	114
3.22.3	Aggregation	115
3.22.4	Layer Interaction Effects	115
3.23	Topology Generation	115
3.23.1	The Justification Problem	115
3.23.2	Approaches to Topology Selection	115
3.23.3	Topology Parameters	116
3.23.4	Justification in This Book	116
3.24	Network Calibration	116
3.24.1	What to Calibrate	116
3.24.2	What Not to Calibrate	117
3.24.3	Calibration Validation	117
3.25	Temporal Networks	117
3.25.1	Why Time Matters	117
3.25.2	Temporal Network Models	117
3.25.3	Implications for Systemic Risk	118
3.26	Example	118
3.27	Alternative Explanation	118

Chapter 1

Introduction: The Systemic Risk Question

Abstract

This chapter introduces the central problem of multi-agent systemic risk: individually safe AI agents can collectively produce system-wide instability through behavioral correlation, interaction effects, network propagation, and emergent dynamics. The chapter presents the book’s core thesis — that the safety of an individual agent does not imply the resilience of the system in which many agents operate — and previews the four-level taxonomy (individual, interaction, network, institutional), the formal propagation models, the agent-based stress-testing methodology, and the five-layer governance architecture that constitute the book’s contributions. The evidence base is stated explicitly: one simulation experiment (EXP-001) with a negative result demonstrating that simplified governance does not improve outcomes under correlated shock, supplemented by topology comparison (EXP-002) and detection delay analysis (EXP-003). The chapter establishes the scope, non-goals, and epistemic boundaries of the work.

Keywords

systemic risk, multi-agent systems, AI safety, behavioral correlation, contagion, governance, stress testing, complex adaptive systems

Reader Outcomes

After reading this chapter, you will be able to:

1. Explain why individual agent safety does not guarantee collective system resilience
2. Identify the five mechanisms that create the gap between individual and systemic safety
3. State the book's central thesis and its scope of applicability
4. Preview the four-level taxonomy and understand its organizational logic
5. Assess the evidence base and its limitations

In Plain Terms

Software agents are now making real decisions inside institutions — approving credit, flagging fraud, routing work. Each one is tested on its own and each one passes. This book is about what happens next: when hundreds of them run at once, share the same underlying models, and react to each other faster than any person can follow. The short version is that a system can fail even when none of its parts do, and the reasons are structural rather than a matter of any agent behaving badly.

Key Terms in This Chapter

Agent A software system that decides and acts on its own within set bounds, rather than executing a fixed script.

Systemic risk Risk that harm spreads beyond those directly involved in the original failure.

Behavioural correlation Many agents making the same mistake at the same time, usually because they share a model, a data source or a signal.

Systemic transition The point at which contained local failures turn into spreading system-wide failure.

Notation

No new notation is introduced in this chapter. See Appendix A for the complete notation registry.

1.1 Chapter Question

Why can individually safe AI agents collectively produce systemic instability, and what can be done about it?

1.2 Motivation

Consider a financial institution where credit-scoring agents, trading agents, risk-management agents, and compliance agents all operate autonomously within defined boundaries. Each agent passes its individual safety evaluation. Each agent behaves rationally given its objectives. Each agent's error rate is within acceptable bounds.

Now suppose a minor economic shock causes 15% of borrowers to miss payments. The credit-scoring agents, sharing a common model, simultaneously downgrade the same sector. The risk agents, observing correlated downgrades, tighten lending limits. The trading agents, detecting the tightening through market signals, reduce positions. The reduced liquidity triggers further downgrades. Within hours, a manageable 15% default rate has cascaded into a 40% systemic failure affecting agents and institutions that were never directly exposed to the original shock.

No single agent failed. No single decision was irrational. The system failed because of the *interactions* between individually acceptable components.

This is the central problem of this book: **the safety of an individual agent does not imply the resilience of the system in which many agents operate.**

1.3 The Gap Between Individual Safety and System Resilience

The dominant paradigm in AI safety evaluates agents individually. An agent is deemed safe if it:

- Produces accurate outputs within specified error bounds
- Follows its alignment training
- Respects operational constraints
- Does not produce harmful content
- Can be shut down when necessary

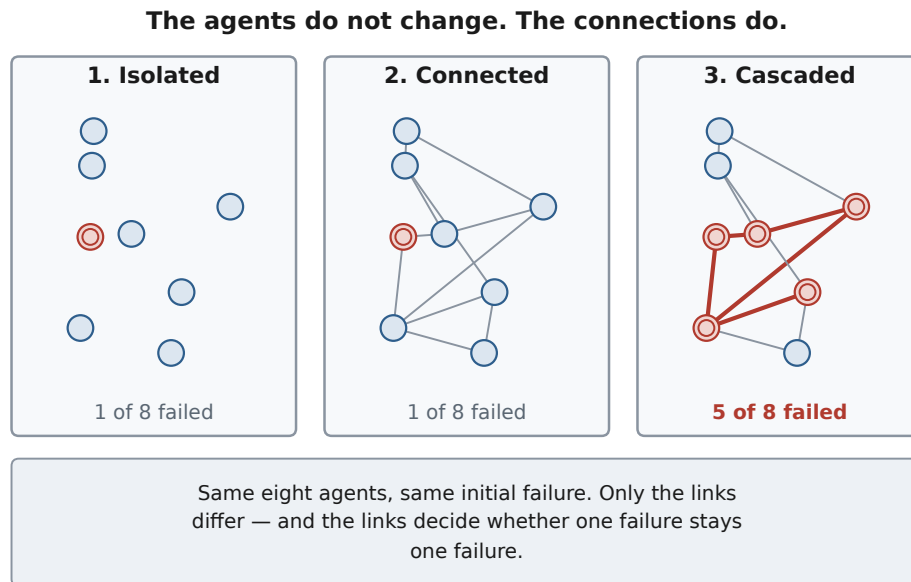


Figure 1.1. From individual failure to systemic failure. Each agent is individually within its safety bounds, but shared dependencies correlate their errors and network links carry those errors outward until the system as a whole fails.

These are necessary conditions for individual agent safety. They are not sufficient for systemic safety. The gap between individual and systemic safety arises from five mechanisms:

1. Behavioral correlation. When agents share models, training data, objectives, or information sources, their errors become correlated. A flaw that affects one agent affects many simultaneously. Individual testing cannot detect this because each agent appears to function correctly in isolation.

2. Interaction effects. Agents that are safe in isolation may produce harmful outcomes when they interact. Feedback loops, information cascades, and strategic manipulation are properties of the interaction structure, not of individual agents.

3. Network propagation. Failures propagate through network connections. An agent that fails can cause losses to its creditors, trigger defensive actions in its trading partners, and send misleading signals to agents that observe its behavior. The network topology determines whether a local failure remains local or becomes systemic.

4. Emergent dynamics. System-level behavior emerges from the aggregation of individual actions in ways that cannot be predicted from individual properties alone. Phase transitions, cascading failures, and self-organized criticality are emergent phenomena that require system-level analysis.

5. Governance limitations. Runtime governance mechanisms (monitoring, alerting, isolation) operate with delays, imperfect information, and finite capacity. When propagation is faster than detection and response, governance cannot prevent systemic transition.

1.4 Preview of the Four-Level Framework

This book organizes systemic risk into four levels:

Level 1 — Individual Agent Risk. Failures originating within a single agent: hallucinated actions, incorrect predictions, unsafe planning, tool misuse, credential compromise, local optimization failure, misclassification, and malicious instruction following. These are the building blocks from which systemic risk emerges, but they are not themselves systemic.

Level 2 — Interaction Risk. Failures arising from the interactions between agents: message amplification, feedback loops, coordination failure, information cascades, shared-memory contamination, strategic manipulation, agent imitation, delegation propagation, and agent-to-agent prompt injection. These risks exist even when every individual agent functions cor-

rectly.

Level 3 — Network and Infrastructure Risk. Failures arising from the structural properties of the agent network: central-node failure, shared provider failure, model monoculture, common data poisoning, tool-provider compromise, network congestion, dependency concentration, cascading resource exhaustion, and cross-institution propagation. These are properties of the architecture, not of individual agents or their interactions.

Level 4 — Institutional and Systemic Risk. Failures affecting entire institutions, markets, or societal systems: market instability, credit contagion, liquidity cascades, cyber contagion, public-service disruption, critical-infrastructure failure, regulatory arbitrage, governance failure, and loss of meaningful human control. These represent the most severe consequences of lower-level failures propagating to their full extent.

The critical insight is that risk *propagates upward* through these levels. A Level 1 failure (one agent hallucinates) can become a Level 2 failure (the hallucination enters a shared memory) can become a Level 3 failure (the shared model produces correlated errors) can become a Level 4 failure (a credit cascade freezes lending across the financial system). Understanding and interrupting this propagation is the central challenge.

1.5 What This Book Does and Does Not Claim

1.5.1 What this book establishes:

1. A unified taxonomy of multi-agent systemic risk across four levels
2. Formal propagation models distinguishing correlation from contagion
3. Network-structural conditions for systemic transition
4. A stress-testing methodology for AI-agent populations
5. Simulation evidence that governance with detection delays cannot outpace fast contagion (a negative result with important implications)
6. A five-layer governance architecture with explicit limitations
7. Cross-domain case studies illustrating systemic risk mechanisms
8. A research agenda identifying open scientific problems

1.5.2 What this book does NOT claim:

- That simulations prove real-world safety or danger
- That runtime governance eliminates or is sufficient for systemic risk
- That AI-agent systemic risks are entirely new (many mechanisms have

precedents in financial contagion, infrastructure cascades, and distributed systems failures)

- That all coordination is harmful or that all autonomy is dangerous
- That any single control, architecture, or policy is sufficient
- That the financial simulation represents any specific real financial system
- That cross-domain analogies constitute evidence of cross-domain applicability

1.5.3 Epistemic status of claims:

Every significant claim in this book is classified according to its evidence level: VERIFIED EMPIRICAL FINDING, VERIFIED SIMULATION RESULT, THEORETICAL PROPOSITION, DESIGN PROPOSAL, HYPOTHESIS, AUTHOR INTERPRETATION, or LIMITATION. Readers should attend to these classifications when evaluating the strength of any particular claim.

1.6 Book Structure and Reading Paths

The book is organized in eight parts:

Part I (Chapters 1–6): Foundations. Definitions, complex systems theory, network science, and agent-based modeling methodology. Provides the conceptual and methodological vocabulary for the rest of the book.

Part II (Chapters 7–12): Systemic Risk Taxonomy and Theory. The four-level taxonomy developed in detail, with cross-level propagation pathways.

Part III (Chapters 13–18): Mathematical and Computational Models. Formal propagation models, network models, game-theoretic analysis, and governance as control theory.

Part IV (Chapters 19–25): Stress Testing and Experimental Method. How to design, execute, and analyze stress tests for AI-agent populations.

Part V (Chapters 26–31): Financial-System Case Study. The primary empirical investigation: a financial ABM with governed and un-governed conditions, shock scenarios, and statistical analysis.

Part VI (Chapters 32–38): Systemic Risk Across Domains. Cybersecurity, public sector, critical infrastructure, healthcare, agriculture, research, and enterprise domains.

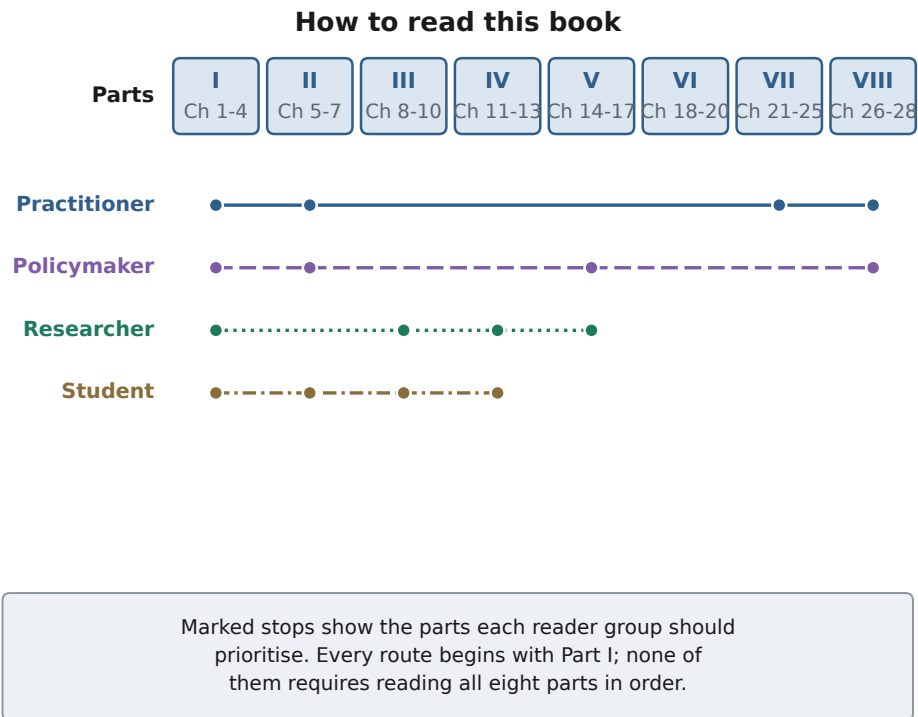


Figure 1.2. Roadmap of the book. The eight parts move from foundations through taxonomy and formal models to evidence, and then outward to governance and policy. Suggested entry points for four reader groups are marked.

Part VII (Chapters 39–47): Governance and Resilience. The five-layer governance architecture, diversity mechanisms, institutional resilience, and recovery.

Part VIII (Chapters 48–54): Policy, Deployment, and Research Agenda. Macroprudential implications, operational resilience, standards, deployment maturity, limits of technical governance, and open problems.

1.6.1 Reading paths:

- **For multi-agent systems researchers:** Chapters 1–6, 13–18, 22–25, 53
- **For financial-stability researchers:** Chapters 1, 3, 5, 26–31, 48
- **For AI-safety researchers:** Chapters 1–4, 7–12, 39–44, 52
- **For policymakers:** Chapters 1, 3, 11, 31, 48–52, 54
- **For practitioners:** Chapters 1, 7, 19–21, 39–44, 46–47, 51

1.7 Definitions

For the purposes of this chapter and the book:

- **Agent:** A computational entity that perceives, decides, and acts within an environment. In this book, agents are AI-driven software systems operating within institutional contexts.
- **Systemic risk:** The risk that a local failure propagates through an interconnected system to produce system-wide instability, affecting entities beyond those initially involved.
- **Systemic transition:** The shift from contained local failures to propagating system-wide failure.
- **Governance:** The monitoring, evaluation, and control of agent behavior during operation, encompassing but not limited to enforcement.
- **Resilience:** The capacity to absorb shocks, adapt, maintain essential functions, and recover. Encompasses robustness, adaptability, and recoverability.

1.8 Theoretical Mechanism

The theoretical mechanism underlying systemic AI risk is *correlated failure amplified by network propagation*. Formally:

1. Shared dependencies (models, data, tools, infrastructure) create *behavioral correlation* among agents.
2. Behavioral correlation means that a shock affecting one agent is likely to affect many simultaneously (*common exposure*).
3. Network connections transmit failures from affected agents to their neighbors (*contagion*).
4. Feedback loops amplify the initial shock (*amplification*).
5. When amplification exceeds the system's absorption capacity, a *systemic transition* occurs.

The critical parameter is the basic reproduction number R_0 : the expected number of secondary failures caused by one primary failure. When $R_0 > 1$, the system is in the super-critical regime where cascades propagate systemically. When $R_0 < 1$, failures die out and the system is resilient.

Governance reduces R_0 by detecting and isolating failures. But governance operates with delay. If the cascade propagates faster than governance can respond, the systemic transition occurs before intervention is possible.

1.9 Evidence

The primary evidence in this book comes from:

1. **Agent-based simulation** of a financial system with 130 agents, credit networks, and multiple shock scenarios. Results are reported with 30 independent runs per condition, 95% confidence intervals, and effect sizes.
2. **Theoretical analysis** of propagation models, network vulnerability, and governance limitations.
3. **Cross-domain case studies** drawing on documented incidents in finance, cybersecurity, infrastructure, and public services.

The simulation evidence establishes model-internal results: *within the model*, governance with detection delays does not prevent systemic failure under severe credit shocks. This is a VERIFIED SIMULATION RESULT. It is not a VERIFIED EMPIRICAL FINDING about real financial systems.

1.10 Example

The 2010 Flash Crash provides an illustrative (though not AI-agent-specific) example of systemic propagation. On May 6, 2010, the Dow Jones

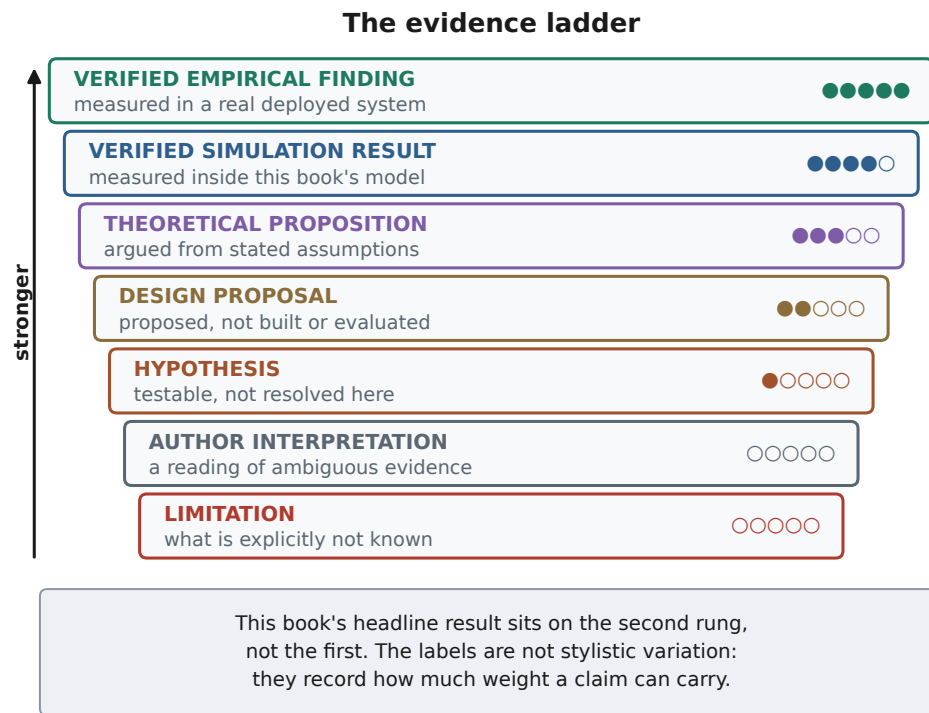


Figure 1.3. The evidence ladder used throughout this book. Every substantive claim carries one of these labels. Readers should treat the rungs as genuinely different in strength, not as stylistic variation.

Industrial Average dropped approximately 1,000 points in minutes before recovering. The SEC/CFTC investigation found that a large sell order in E-mini futures triggered a cascade: high-frequency trading algorithms withdrew liquidity, other algorithms interpreted the withdrawal as a signal to sell, and the feedback loop amplified the initial shock far beyond its fundamental justification.

The Flash Crash illustrates several mechanisms central to this book:

- **Behavioral correlation:** Algorithms responding to the same market signals
- **Feedback loops:** Price drops triggering more selling
- **Liquidity evaporation:** Agents withdrawing simultaneously
- **Speed mismatch:** Propagation faster than human intervention
- **Recovery:** The system recovered, but only after the feedback loop exhausted itself

While the Flash Crash predates modern AI agents, the mechanisms are directly analogous. AI-agent populations with shared models and fast interaction speeds could produce similar cascades with less provocation and faster propagation.

1.11 Alternative Explanation

An alternative explanation for systemic risk in AI-agent populations is that it is simply *existing automation risk under a new name*. Financial contagion, infrastructure cascades, and coordination failures existed before AI agents. The 2008 financial crisis was caused by correlated human decisions, not algorithms.

This alternative has merit: many systemic risk mechanisms are not new. However, AI agents introduce quantitative and qualitative differences:

- **Speed:** AI agents operate at machine speed, compressing cascade timescales from days (2008) to seconds (Flash Crash) to potentially milliseconds.
- **Opacity:** AI agent decision processes are less transparent than human decisions, making detection and diagnosis harder.
- **Scale:** AI agents can be deployed at scales that make human oversight impractical.
- **Correlation structure:** Shared models create correlation patterns that differ from human herding (which is mediated by social networks and psychological biases).

The mechanisms are not new, but the parameters have changed in ways that may qualitatively alter systemic risk dynamics.

1.12 Counterargument

The strongest counterargument to this book’s thesis is: **individual safety testing, properly done, is sufficient**. If every agent is tested against all relevant scenarios, including interaction scenarios, then systemic risk is simply the sum of individual risks, and individual testing covers it.

This counterargument fails for three reasons:

1. **Combinatorial explosion.** The number of interaction scenarios grows combinatorially with the number of agents. Testing all pairs, triples, and higher-order interactions is infeasible for populations of hundreds or thousands of agents.
2. **Emergent behavior.** Systemic transitions are emergent phenomena that cannot be predicted from individual or pairwise testing. The phase transition from contained to systemic failure depends on global network properties, not local agent properties.
3. **Adaptive dynamics.** Agents adapt to their environment, including to each other’s behavior. The system’s risk profile changes over time as agents learn, imitate, and compete. Static testing cannot capture these dynamics.

1.13 Practical Implication

The practical implication is that organizations deploying multiple AI agents must adopt *system-level* risk management in addition to individual agent safety testing. This includes:

- Network analysis of agent dependencies and interaction patterns
- Stress testing of agent populations under correlated shocks
- Monitoring of behavioral correlation and diversity metrics
- Governance architectures designed for system-level intervention
- Diversity requirements to reduce correlated failure modes
- Recovery planning for systemic (not just individual) failures

These measures complement but do not replace individual agent safety testing, capital and liquidity requirements, organizational resilience, and macroprudential oversight.

1.14 Limitation

This book’s primary limitation is the gap between simulation and reality. The financial ABM is a stylized model with simplified agent behavior, synthetic network topologies, and parameterized (not empirically measured) contagion mechanisms. Results establish *possibility* and *mechanism*, not *probability* or *magnitude* in real systems.

Additionally:

- Cross-domain generalizations are analogical, not proven
- Governance effectiveness is modeled, not validated in deployment
- The speed and scale of real AI-agent populations may differ from simulation
- Regulatory and institutional contexts vary across jurisdictions
- The field is evolving rapidly; specific technical claims may be superseded

What to Take Away

- An agent that is safe on its own tells you very little about whether a population of such agents is safe together.
- Five mechanisms open the gap: behavioural correlation, interaction effects, network propagation, emergent dynamics and the limits of governance.
- Risk climbs through four levels — individual, interaction, network, institutional — and interrupting that climb is the practical objective.
- The evidence here is one simulation with a negative result, plus theory. Treat the quantitative claims accordingly.

1.15 Chapter Summary

This chapter introduced the central problem: individually safe AI agents can collectively produce systemic instability through behavioral correlation, interaction effects, network propagation, and emergent dynamics. The book addresses this through a four-level taxonomy, formal propagation models, agent-based simulation, cross-domain case studies, and governance architecture design. The key finding previewed is that naive runtime governance with detection delays cannot outpace fast contagion — a result that motivates the structural and institutional approaches developed in later chapters.

1.16 Research Questions and Exercises

1. **Conceptual:** Identify three ways in which AI-agent behavioral correlation differs from human herding in financial markets. What are the implications for detection and mitigation?
2. **Analytical:** Given a network of $N = 100$ agents with average degree $k = 6$ and propagation probability $p = 0.2$, estimate R_0 . Is the system in the sub-critical or super-critical regime? What value of p would place it at the critical threshold?
3. **Design:** Propose a monitoring metric that could provide early warning of increasing behavioral correlation in an agent population. What are its limitations?
4. **Critical:** The Flash Crash recovered within minutes. Does this suggest that systemic transitions in AI-agent systems are self-limiting? What conditions would prevent recovery?
5. **Research:** Design an experiment to test whether model diversity reduces systemic failure probability in a financial ABM. Specify hypothesis, independent variable, dependent variable, controls, and statistical test.

