Nobody Planned This

Artificial Superintelligence as an Emergent Phenomenon

Nils Baierl

March 2026

Abstract

This essay argues that Artificial Superintelligence (ASI) will not emerge through parameter scaling alone, but through the emergent properties of multi-agent systems. Three years after Google DeepMind's "Levels of AGI" framework classified current systems as Level 1 (Emerging AGI), we stand at an inflection point. Drawing on Marvin Minsky's Society of Mind, Langton's Ant, and recent evidence from METR's 2026 capability horizon study, I propose that the path to ASI lies in *topology scaling* — the combinatorial expansion of agent interactions — rather than monolithic model growth. Intelligence, at its highest levels, is collective. The highway is forming. Nobody planned it.

1 Introduction: The Shifting Baseline

Claims that "AGI is already here" have become ubiquitous. They are not entirely wrong.

In 2023, Google DeepMind published a framework for classifying AGI progress[1]. Their taxonomy defined six levels based on performance, generality, and autonomy. At publication, current systems — GPT-4, Claude, Gemini — were classified as **Level 1: Emerging AGI**, defined as "equal to or somewhat better than an unskilled human."

We are now three years later. The baseline has shifted.

The question is no longer whether AGI is "here." The question is: what emerges from it?

This essay proposes a specific answer: Artificial Superintelligence will not arise through continued scaling of monolithic models. It will *emerge* from the combinatorial dynamics of multi-agent systems. The mechanism is not bigger models. It is *topology scaling* — the expansion of agent interactions across domains, hierarchies, and time. Intelligence, at its highest levels, is collective.

2 Levels of AGI: Three Years Later

DeepMind's 2023 framework classified current systems as **Level 1: Emerging AGI** — capable at the level of an unskilled human, but with uneven performance across domains.

[1] Levels of AGI: Operationalizing Progress on the Path to AGI, Morris et al., DeepMind, 2023

Figure 1 illustrates this first phase transition: decades of narrow AI gave way to foundation models with emergent capabilities (reasoning, in-context learning, code generation) through transformer architecture and parameter scaling. This was the first emergence. It brought us to Level 1.

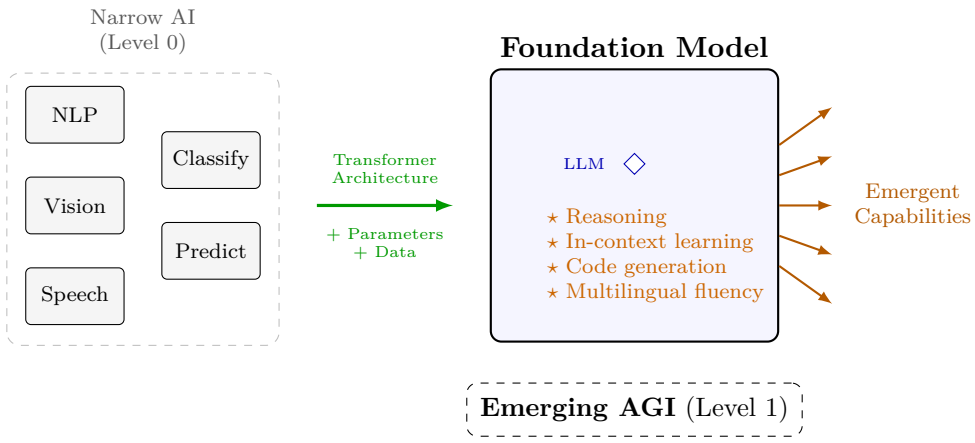

Figure 1: The first emergence. Decades of narrow, specialized AI models (Level 0) gave way to foundation models through transformer architecture and massive parameter scaling. The result: emergent capabilities — reasoning, in-context learning, code generation — that define Level 1 (Emerging AGI). A quantitative change in scale produced a qualitative change in kind.

Three years on, the landscape has shifted.

A 2026 study by METR (Model Evaluation & Threat Research) measured AI capability using a new metric: *50%-task-completion time horizon* — the duration of tasks that AI models can complete with 50% success rate[2]. The study evaluated 12 frontier models (2019–2025) on 170 software engineering tasks, with human baselines from 800+ skilled professionals totaling over 2,500 hours.

Results: Capability has been **doubling every seven months** since 2019, with the rate accelerating since 2023. Extrapolation predicts AI systems will reach a one-month time horizon between 2028 and 2031.

The implication: static classifications like DeepMind's levels capture a moment in time, but the trajectory is what matters. A seven-month doubling time means the baseline has shifted substantially since 2023. The first emergence — the arrival of Level 1 — was only the beginning.

3 Society of Mind: Intelligence as Multiplicity

Marvin Minsky's *Society of Mind* (1986) offers a different origin story for intelligence. The central claim is radical in its simplicity: intelligence does not arise from a single, unified process. It emerges from the interaction of many simple, unintelligent agents.

> "What magical trick makes us intelligent? The trick is that there is no trick. The power of intelligence stems from our vast diversity, not from any single,

[2]Measuring AI Ability to Complete Long Software Tasks, arXiv:2503.14499v3, Kwa et al., Feb 2026

perfect principle."

Minsky's agents are not LLMs. They are minimal processes — each capable of one trivial task. An agent that recognizes vertical lines. An agent that detects motion. An agent that compares two quantities. None are intelligent on their own. Together, they constitute what we experience as thought.

This is not metaphor. It is architecture.

4 Langton's Ant: Emergence in Action

Consider Langton's Ant, a cellular automaton devised by Chris Langton in 1986. The rules are trivial:

1. Place an "ant" on a grid of white cells.
2. If the ant is on a white cell, turn right, flip the cell to black, move forward.
3. If the ant is on a black cell, turn left, flip the cell to white, move forward.
4. Repeat.

For thousands of steps, the ant's path appears chaotic. No pattern is visible. Then, abruptly, the system transitions into what is called a "highway": a diagonal, infinitely repeating sequence that continues until the grid boundary is reached.

Nobody designed the highway. It was not programmed. It *emerged* from the interaction of simple rules applied iteratively.

This is **emergence**: the phenomenon wherein new, higher-level properties arise in hierarchically organized systems that cannot be predicted from the properties of the interacting elements at lower levels, but instead appear in unforeseen ways.[3] In computer science, emergence is defined more operationally as the occurrence of novel qualities during state changes in a system.[4]

As Aristotle observed: "The whole is more than the sum of its parts" (*to holon para ta mera*) — an early articulation of emergentism.[5]

Multi-agent systems operate on the same principle. Individual agents follow local rules. Their interactions, at scale, produce global behavior — planning, reasoning, coordination — that no single agent possesses.

5 Evidence from Multi-Agent Systems

The claim that multi-agent interaction produces emergent capabilities is not merely theoretical. Empirical evidence has accumulated over the past decade.

[3]Cf. Wiktionary: "emergence" — the phenomenon of new, higher-level properties arising in organized systems unpredictably from lower-level interactions.

[4]Ibid., computer science sense: novel system properties arising from state transitions.

[5]Aristotle, *Metaphysics* 1041b11; often cited as a foundational statement of emergence: the whole possesses properties not present in its constituent parts.

In 2019, OpenAI demonstrated that agents playing a simple game of hide-and-seek developed *tool use* without explicit programming.[6] Through multi-agent competition and self-play, agents discovered six distinct strategic phases: hiding, shelter-building, ramp usage, and door-locking. None of these behaviors were hardcoded. They *emerged* from the autocurriculum created by adversarial interaction.

DeepMind observed similar phenomena in sequential social dilemmas[7] and in Capture the Flag gameplay[8], where agents learned complex team coordination, human-competitive strategies, and tool use through self-play. These systems demonstrated that multi-agent interaction can produce capabilities that exceed the design of individual agents.

More recently, studies of LLM-based multi-agent systems have confirmed these patterns. A 2025 study from the University of Wisconsin-Madison found that multi-agent LLM systems exhibit *genuine emergent properties* — group-level dynamics not reducible to individual behaviors.[9] Even modest models (e.g., Llama) showed temporal synergy when arranged in multi-agent topologies. The MAEBE framework (2025) systematically measured emergent behavior across different group structures (ring, star, heterogeneous) and found that *topology itself* determines the nature and extent of emergence.[10]

Quantitative analyses corroborate these findings. A 2025 comparative study reported that multi-agent systems exhibit **37.2% improved reliability** and **22.8% better zero-shot generalization** compared to single-agent baselines.[11]

These results do not prove that multi-agent systems will inevitably produce ASI. But they establish a clear pattern: interaction topology is a distinct axis of scaling, orthogonal to parameter count, that produces emergent capabilities.

6 Scaling as Topology

The dominant narrative of AI progress is scaling: more parameters, more data, more compute. This is not wrong, but it is incomplete. The first emergence (Figure 1) was driven by exactly this: parameter scaling. But the *second* emergence will not come from making models bigger. It will come from making them *many*.

Scaling matters, but what is being scaled is changing. It is no longer only model size. It is *topology* — the structure of interactions between agents. Figure 2 illustrates this:

A multi-agent system scales not through parameters, but through *topology*. What changes is not the capability of any single agent, but the structure of their interaction. The system becomes more capable through orchestration — how agents divide labor, share context, and coordinate action. How exactly this orchestration works remains an open question; current multi-agent systems are early prototypes, not a final architecture. What matters is the direction of travel: from single agents to many, from isolation to interaction.[12]

[6] Emergent Tool Use From Multi-Agent Autocurricula, Baker et al., OpenAI, 2019 (arXiv:1909.07528).

[7] Multi-Agent Reinforcement Learning in Sequential Social Dilemmas, Leibo et al., DeepMind, 2017.

[8] Human-level performance in a 3D multiplayer game with population-based reinforcement learning, Jaderberg et al., DeepMind, 2019.

[9] Emergent Coordination in Multi-Agent Language Models, Hodima, 2025.

[10] MAEBE: Multi-Agent Emergent Behavior Framework, 2025 (arXiv:2506.03053).

[11] Multi-agent systems: the future of distributed AI platforms for complex problems, WJARR, 2025.

[12] Multi-agent systems are not a new invention — research into them predates the current LLM era

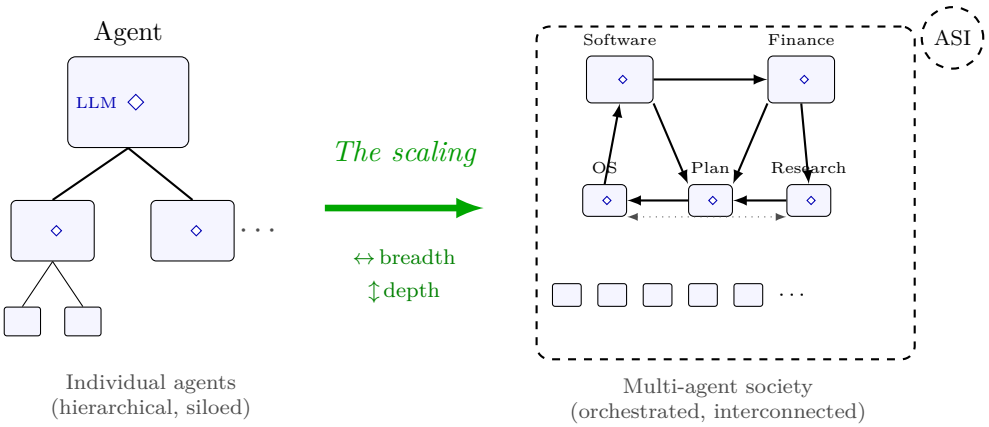

Figure 2: The second emergence: topology scaling. *Left*: individual LLM-powered agents in simple hierarchies. *Center*: the axis of scaling shifts — not just more parameters, but more *topology*: breadth (more domains) and depth (richer hierarchies). *Right*: A resulting multi-agent society, densely interconnected.

This is the argument in two acts. The first emergence (Figure 1) was parameter scaling producing foundation models with capabilities no narrow system had. The second emergence, I propose, will be topology scaling producing collective intelligence that no individual model has.

Hypothesis: ASI arises not from the scaling of a single agent, but from the scaling of *many agents in interaction*. The exact mechanisms of this emergence — the specific orchestration patterns, communication protocols, and coordination structures that maximize collective capability — remain an open research question. What is clear is the direction: from single agents to many, from isolation to interaction.

This hypothesis is grounded in three precedents. First, Minsky's Society of Mind offers a cognitive blueprint: human intelligence itself may be the product of interacting subagents. Second, Langton's Ant demonstrates that simple interaction rules can produce unpredictable, higher-order structure. Third, emergent systems like the internet and the global economy show that unplanned, decentralized interaction can generate functional complexity at planetary scale.

None of these precedents guarantee that multi-agent AI will produce ASI. But they establish that emergence from interaction is not merely possible — it is a recurring pattern in complex systems. Minsky's society is not static. It grows. And in that growth, new capabilities may appear that were not present in any constituent part.

7 Caveats and Outlook

The argument presented here is a *plausibility case*, not a proof. That multi-agent systems exhibit emergent behavior is empirically established. That this emergence will inevitably lead to ASI is not.

by decades, with foundational work emerging in the 1980s and 1990s alongside early distributed AI and robotics. The paradigm shift is not their existence, but the scale and quality at which we can now orchestrate them. It is this orchestration — applied at sufficient breadth and depth — that creates the conditions for emergent superintelligence.

Three important caveats apply. First, most evidence for multi-agent emergence comes from *game environments* (hide-and-seek, Capture the Flag, social dilemmas) rather than open-ended real-world tasks. Whether the same dynamics generalize to software engineering, scientific discovery, or strategic planning remains an open empirical question.

Second, emergence is a *descriptive* concept, not a *predictive* one. We can recognize emergence after it occurs; we cannot reliably predict *what* will emerge from a given topology. This is both the promise and the risk of multi-agent systems: they may produce capabilities we did not design — for better or worse.

Third, the leap from "emergent coordination" to "superintelligence" is substantial. The evidence reviewed here shows that multi-agent systems can exceed individual agent capabilities. Whether they can *recursively* improve themselves, develop open-ended generality, or achieve the kind of strategic reasoning associated with ASI is not demonstrated by current work. It is a *hypothesis* — grounded in precedent, but unproven.

With these caveats in mind, I maintain the core claim: if ASI emerges, it will likely do so through multi-agent topology scaling rather than monolithic parameter scaling. The precedents — Minsky's Society of Mind, Langton's Ant, empirical multi-agent results — point in this direction. The accelerating capability horizon documented by METR suggests the timeline may be short. Whether this trajectory leads to ASI, to powerful but bounded systems, or to unexpected failure modes is a question for future work.

What is clear is that the axis of scaling has shifted. The next decade of AI progress will be defined not by how large we can make models, but by how well we can orchestrate many models in interaction. The highway is forming. Nobody planned it. But we may still be able to steer it.

8 Conclusion

Nobody planned the highway in Langton's Ant. Nobody planned the internet. Nobody planned the global economy. These are emergent systems — order arising from interaction, not design.

ASI will be no different. It will not arrive as a single, finished artifact. It will emerge from the scaling of multi-agent systems, from the collective dynamics of many models interacting, competing, and cooperating.

The question is not whether this will happen. It is whether we will understand it while there is still time to shape it.