

Approaching AGI and ASI with Network Intelligence

Vision Paper by Tomas Bengtsson, IXO
December 2025

Abstract

We hypothesize that continued scaling of large language models (LLMs) alone is approaching diminishing returns, and that further progress toward artificial general intelligence (AGI) will primarily arise from advances in algorithmic engineering, curation, and system-level architecture. Beyond AGI, additional research will be required to enable artificial superintelligence (ASI).

As millions of models and agents are being deployed across domains, a clear trend is emerging: increasingly specialized agents are being developed for all narrow tasks, tools, and environments. The central challenge is no longer the creation of individual capabilities, but enabling these heterogeneous agents to work together. If such agents can be dynamically coordinated and composed, the resulting system can perform the full range of human cognitive tasks, consistent with a functional realization of AGI.

We propose an adaptive network architecture, termed Network Intelligence, in which large numbers of specialized models and agents are composed into a unified cognitive system. We introduce the concept of a Composable Agent Network, supported by open protocols analogous to the Internet, enabling decentralized development, interoperability, and community-driven scalability. Core mechanisms such as planning, model routing, and adaptive orchestration allow the system to discover, select, combine, and create agents on demand.

Extending this framework, we argue that algorithms capable of intelligently composing and reconfiguring the agent network at runtime can enable continuous learning at the system level. As foundation models improve, their capabilities are leveraged across the network. Coupled with a continuously evolving world model, such adaptive network learning provides a plausible pathway from AGI toward emergent ASI.

1. Introduction

Early work on this topic was motivated by concerns regarding the long-term **existential risks** posed by the development of artificial superintelligence (ASI). In particular, the question was whether increasingly capable AI systems could be pursued along pathways that reduce the probability of irreversible or catastrophic outcomes for humanity. Initial efforts focused on alignment and control research; however, this line of inquiry did not yield a clear or sufficient framework for ensuring safety in the presence of highly autonomous, capable, and adaptive systems.

This motivated a shift in perspective. Rather than addressing ASI solely as a post-hoc control or alignment problem, this work explores whether properties of the system architecture itself, such as decentralization, openness, and adaptive composition, can influence risk profiles at a fundamental level. The central hypothesis is that certain architectural choices may reduce concentration of power, limit single-point failures, and constrain runaway dynamics, thereby mitigating classes of existential risk.

Accordingly, this paper presents a hypothesis on a possible route toward a **decentralized and adaptive form of intelligence**. The proposed approach emphasizes network-level composition of specialized agents, open protocols, and continuous reconfiguration. While strong claims about objectivity or global “truth-seeking” are necessarily limited, the goal is to explore whether such architectures can reduce systematic bias and fragility relative to centralized monolithic systems, and thereby offer a potentially safer trajectory toward advanced AI.

1.1 Defining Progress Toward AGI and ASI

The paper **Levels of AGI: Operationalizing Progress on the Path to AGI** (Google DeepMind) proposes a framework for characterizing progress toward AGI along three dimensions: performance, generality, and autonomy. Within this framework, AGI is reached when an AI system matches or exceeds the performance of a high percentile of skilled adults across a broad range of tasks, while artificial superintelligence (ASI) corresponds to performance exceeding that of all humans across essentially all tasks.

Building on this perspective, this paper adopts a pragmatic view of evaluation. In real-world settings, the most comprehensive and continuously exercised mechanism for assessing value creation is the economic system. Accordingly, we hypothesize that a necessary—though not sufficient—criterion for viable AGI or ASI is the ability to generate sustained economic value across a wide range of tasks and domains.

Under this view, progress toward AGI can be operationalized not only through benchmark performance, but also through demonstrated **economic usefulness**, adaptability, and autonomy in real-world task environments. This framing does not claim that economic value fully captures intelligence or safety, but rather that it provides a scalable, externally validated signal of general capability and practical impact.

1.2 Network Intelligence

As models and agents are increasingly developed for a wide range of human tasks, it becomes feasible to connect such components into a networked system. In this paper, we use the term models to collectively refer to both models and agents, unless otherwise specified. Within an appropriate architectural framework, individual models, or compositions of multiple models, can be dynamically selected to address specific queries or tasks.

A network-based approach offers several potential advantages. First, heterogeneous architectures can coexist within the same system, allowing models with different training objectives, modalities, and inductive biases to be combined. Rather than relying on a single monolithic generalist, the network can incorporate highly specialized models optimized for particular tasks. Prior work on multi-agent systems suggests that such specialization and coordination can improve overall system performance, **analogous to role differentiation within an organization**.

Beyond coordination, the network itself can be treated as an adaptive computational structure. Interactions between models can be represented as **weighted connections**, which may be strengthened or weakened over time based on performance, feedback, or contextual relevance. This perspective allows the system to be viewed at a higher level of abstraction, analogous to a neural network operating over models rather than parameters.

Additional mechanisms such as planning, task decomposition, and adaptive routing can further enhance system-level capability. Together, these components enable runtime adaptation and **learning at the network level**, providing a foundation for increasing intelligence through composition rather than solely through model scaling.

2. Empirical Trends and Motivation

The predictions and hypotheses explored in this paper are derived primarily from first-principles reasoning, and are further informed by two observable developments: the **large-scale deployment of task-specialized models** and agents, and a growing body of empirical evidence demonstrating the **effectiveness of multi-agent systems** (MAS). The rapid proliferation of specialized models across tasks and domains, together with consistent performance gains from coordinated multi-agent architectures, provides a concrete and empirically grounded motivation for investigating system-level approaches to intelligence.

The deployment of models and agents is accelerating across both open and commercial ecosystems. Public repositories such as Hugging Face host over 2.3 million models, many of which are specialized for narrow tasks, domains, or modalities. In parallel, traditional enterprises increasingly fine-tune, integrate, and deploy models that partially or fully replace specific organizational roles, such as customer support, data analysis, and software development. At a larger scale, industry initiatives further illustrate this trend; for example, Salesforce has publicly stated an ambition to deploy up to one billion agents within its ecosystem by the end of 2025.

These developments suggest a shift in emphasis from improving individual model performance toward coordinating large numbers of heterogeneous, specialized components. We hypothesize that artificial general intelligence may emerge not solely from increasingly capable individual models, but from **effective composition, coordination, and orchestration** of many such models. Under this view, progress toward AGI depends increasingly on “better coordination” rather than exclusively on “better models.”

This hypothesis is supported by a growing body of empirical work demonstrating that **multi-agent systems outperform single large language models** across a variety of tasks. Recent studies show that multi-agent collaboration improves software engineering correctness by approximately 31% and reduces bugs by 23% in structured development pipelines (Qian et al., 2024). Role-specialized agent interactions have been shown to increase multi-hop reasoning and complex coding accuracy by 16–40% (Li et al., 2023), while explicit role decomposition achieves up to 2.7× higher functional correctness in multi-file code generation tasks (Hong et al., 2023). In interactive web-based environments, multi-agent coordination surpasses single-agent baselines by 22–30% on WebArena and yields approximately 25% gains on GAIA and MiniWoB++ (Duan et al., 2024). More recent work indicates that self-collaboration, where agents reason with copies of themselves, improves accuracy by 28–42% on benchmarks such as GSM8K, StrategyQA, and MetaMath (Shen et al., 2025).

Taken together, these results provide converging evidence that multi-agent architectures can deliver systematically higher performance than single-model approaches across reasoning, coding, and interactive task domains. While these findings do not establish a direct path to AGI, they motivate the investigation of networked, compositional intelligence as a plausible and empirically grounded direction for further research.

3. Network Intelligence: Architecture

This section introduces Network Intelligence as an architectural paradigm for composing large numbers of specialized models and agents into a coherent, adaptive system. The focus is on system-level structure and coordination mechanisms rather than individual model capabilities, and on how such architectures may support increasing levels of generality, autonomy, and learning over time.

3.1 Concept Overview

Network Intelligence is defined as an adaptive, network-based architecture in which models and agents are treated as interoperable components that can be dynamically discovered, created, composed, evaluated, and replaced. In this framework, intelligence emerges not solely from the capabilities of individual models, but from the structure and dynamics of the network that connects them.

Models and agents are represented as nodes in a network, where edges encode interaction patterns such as information flow, delegation, verification, or coordination. Rather than assuming a static set of components, the network supports **dynamic composition**, allowing the system to select individual models or combinations of models in response to a given task or goal. As new or improved models become available, existing components can be replaced or deprioritized without requiring global retraining or redesign.

An analogy can be drawn to ranking and discovery mechanisms in large-scale information networks. Just as search engines dynamically surface relevant pages based on evolving signals, a Network Intelligence architecture can prioritize models based on performance, reliability, cost, or contextual relevance. This allows the system to continuously adapt as the underlying model ecosystem evolves.

3.2 Composable Agent Network

A central design principle of Network Intelligence is **composability**. The architecture assumes that models and agents are developed independently and may differ widely in implementation, scale, modality, and purpose. To enable coordination at scale, the system relies on open protocols analogous to those that underpin the Internet. These protocols define how agents describe their capabilities, accept tasks, exchange information, and report outcomes.

This approach enables interoperability and decentralization, reducing dependence on any single provider or model class. Rather than a centrally controlled system, the network can grow organically as new agents are introduced by a broad community of contributors. Such decentralization supports resilience, diversity of approaches, and rapid innovation, while mitigating single points of failure.

Community-driven scaling further allows the network to reflect a wide range of domains and perspectives. As with open software ecosystems, value emerges from the ability to integrate heterogeneous components into larger systems, rather than from uniformity or central optimization alone.

In addition to individual and academic contributors, the architecture allows commercial entities to publish models and agents into the network under clearly defined interfaces and protocols. Companies can expose specialized capabilities, such as domain expertise, proprietary tools, or optimized services, while retaining control over deployment, access conditions, and updates. This enables organizations to participate in a shared ecosystem without relinquishing ownership or competitive differentiation, while still benefiting from interoperability and composability.

To support sustained participation, the architecture can incorporate **incentive and compensation mechanisms** that allow individuals and organizations to monetize the use of their published models and agents. Compensation may be tied to measurable usage, performance, or value contribution within the network, while remaining compatible with open standards and decentralized governance. Such mechanisms are intended to align incentives for quality, maintenance, and continued

improvement, without requiring centralized ownership or exclusive control. Similar to existing platform ecosystems, this allows public and private components to coexist, enabling scalable innovation through contribution rather than centralization.

3.3 Coordination Mechanisms

Effective coordination is critical for transforming a collection of specialized agents into a system with general capability. Network Intelligence relies on several core coordination mechanisms.

Goal setting provides direction at the system level. Goals are defined as desired future world states and serve as reference points for planning and evaluation. These goals may be externally specified or internally generated, and can operate at different levels of abstraction.

Planning enables the system to map paths from the current state to a desired goal state. This includes the ability to decompose objectives, sequence actions, allocate agents, and monitor progress. Planning is treated as a continuous process rather than a one-time computation, incorporating replanning and course correction as new information becomes available. This aligns planning with value and control functions, allowing the system to adjust behavior in response to feedback. As noted by Kahneman, "Intelligence is not only the ability to reason; it is also the ability to find relevant material in memory and to deploy attention when needed." Continuous planning integrates reasoning, attention, prediction, and execution within a unified control loop.

Model routing determines which models or agents are invoked for a given subtask. Routing mechanisms account for factors such as task requirements, model specialization, performance history, cost, and latency. This includes the continuous integration of improved foundation models as well as large numbers of smaller, task-specific models, such as small language models (SLMs). Routing decisions are adaptive and can evolve over time as the network changes.

Adaptive orchestration governs the interaction between agents during task execution. This includes managing dependencies, resolving conflicts, aggregating outputs, and coordinating verification or refinement steps. Orchestration operates across temporal scales, from short-lived task execution to longer-term system optimization.

Task decomposition allows complex objectives to be broken down into smaller, more manageable components that can be handled by specialized agents. Decomposition strategies may be hierarchical, parallel, or iterative, and are closely coupled with planning and orchestration mechanisms.

3.4 Network-Level Learning and Adaptation

To move beyond static coordination toward increasing levels of intelligence, Network Intelligence treats the network itself as a **higher-level computational and learning structure**. At this level, learning occurs not only within individual models, but across their interactions.

Connections between models can be represented as **weighted relationships**, reflecting factors such as trust, relevance, performance, or contextual suitability. These weights can be adjusted over time based on feedback, allowing the network to strengthen effective pathways and weaken less useful ones. Such adaptation may be implemented through engineered heuristics, learned policies, or hybrid approaches.

Runtime adaptation and learning are central to this framework. Rather than relying exclusively on offline training, the network is designed to learn through interaction with tasks and environments. This aligns with the view that advanced intelligence requires continuous learning in open-ended settings. As Pedro Domingos observes, “The key to efficient learning is not the verifiability of the final result; it’s getting feedback at every step.” Network-level learning emphasizes incremental feedback, evaluation, and adjustment during execution.

A critical component of this process is the construction and maintenance of a **world model**. The system must acquire and integrate information about its environment in order to generalize across tasks and contexts. As with human learning, this requires active exploration and interaction, not merely passive observation. A world model is continually updated as new information becomes available, and its quality directly constrains the effectiveness of reasoning and planning. High capability in isolation is insufficient without an accurate and well-grounded model of the world.

Finally, **memory** provides persistence across time. Memory mechanisms store task outcomes, model performance data, environmental information, and learned abstractions. While memory can be implemented through engineering solutions, it may be extended with algorithmic mechanisms such as selective retention, abstraction, and forgetting. Controlled forgetting is particularly important for maintaining relevance and adaptability as the system evolves.

4. Safety and Conclusions

Despite substantial research attention, the alignment and control problem for highly capable AI systems **remains unresolved**. In particular, no generally accepted solution has emerged that can guarantee safe behavior under conditions of high autonomy, open-ended learning, and strategic reasoning. This suggests that alignment may not be solvable solely through post-hoc constraints or centralized control mechanisms.

In current deployments, risks also arise from institutional and economic incentives that shape system behavior, including bias, opacity, and concentration of power. While such factors are not unique to AI systems, they can be amplified as capabilities scale. Approaches that emphasize transparency, pluralism, and mechanisms for cross-validation, such as incorporating multiple independent models and perspectives, may help mitigate some of these risks, though they do not eliminate them.

With respect to existential risk to humanity, uncertainty is unavoidable. The development of increasingly capable AI systems necessarily involves unknown failure modes and emergent behaviors. Consequently, this work does not claim to offer definitive safety guarantees. Instead, it explores whether architectural properties such as decentralization, composability, and adaptive oversight can **reduce specific classes of risk**, including single-point failures and unilateral control, relative to monolithic and centrally governed systems.