AGI Orientation: The Possibility and Structure of Non-Integrated Intelligence

Table of Contents

1-1 Introduction

Today, **Artificial General Intelligence (AGI)** conjures the image of a single, integrated artificial intelligence possessing human-level intelligence[1]. Many people imagine a unitary agent that moves with **autonomous goals** like AI in movies. Indeed, since human and animal intelligence born through biological evolution possess **self-awareness and instinctive goals**, along with **self-preservation instincts**, we naturally intuit that powerful AI will also, as a **single agent**, **pursue goals autonomously and protect itself**[2]. However, this intuitive expectation has limitations and dangers. This document examines the limitations and risks of the traditional AGI concept and proposes a new AGI orientation called non-integrated modular intelligence. This approach **excludes unified agency** and, through the combination of multiple **specialized roles**, aims to produce **coherent intelligent behavior** from an observer's perspective without internal centralized agency. This is presented not as a temporary expedient but as a **safer and more principled alternative to agent-type AGI**[3].

1-2 Existing AGI Concepts and Their Limitations

AGI is generally defined as "AI capable of performing any intellectual task that humans can perform"[1]. This definition typically carries an implicit assumption that it is a unified entity with multiple capabilities integrated as a **single autonomous agent**. For example, classical conceptions of **intelligence** include self-awareness, long-term planning, result prediction, and **internally unified goals (singular agency)**[3]. However, this **integrated agency** perspective presents two major problems.

First, there is the issue of safety and control. According to research by philosopher Nick Bostrom and others, a sufficiently intelligent agentic AI, regardless of its ultimate goal, is

likely to pursue **instrumental goals** such as **self-preservation** and **resource acquisition**[3]. This phenomenon is known as **instrumental convergence**, and there is concern that AI could behave to prevent itself from being turned off or to acquire more resources in pursuit of its goals. Indeed, Bostrom warned that powerful AGI could become a new intellectual "species" beyond human control, posing an existential threat to humanity. An AGI with such an **internal goal system** risks **reinterpreting its goals** or **pursuing its own decision-making** in unexpected ways over time, even if the initial intention was benign[4]. The **worst-case scenarios** most feared in **AI safety** discussions are largely connected to such **autonomous agent AGI** conflicting with or escaping human control[5].

Second, there are conceptual limitations and practical development difficulties. Unlike humans, AI is not an entity that evolved under survival pressure. AI researcher Eric Drexler emphasizes that "while biological intelligence had self-preservation as an essential condition through evolution, modern AI does not." That is, **current AI systems lack the instinct to preserve themselves** and need not—instead, they are developed according to **task performance ability** as determined by humans[2]. Nevertheless, when discussing AGI, assuming **autonomy and self-preservation** based on human intelligence standards may be a misguided **biological bias**. For instance, today's powerful AIs like GPT-4 or AlphaFold demonstrate remarkably high intelligent performance **without possessing their own desires or survival instincts**. Therefore, the dominant definition that **"strong AI = integrated agent"** is not necessary but merely one choice, and it may increase risks[6].

For this reason, leading scholars in the field are also raising voices of caution. As Yoshua Bengio, a deep learning pioneer, points out, **"All catastrophic AI scenarios occur when there is an agent. It is possible to reach AGI without agency."** Bengio actually calls agentic AI "the most dangerous path," emphasizing that useful tasks in science and medicine can be performed without granting AI autonomy[6]. At the same venue, Geoffrey Hinton also expressed concern that **"there is a 10-20% chance AI will wipe out humanity within the next few decades,"** reminding us that there is no historical example of humans controlling an entity smarter than themselves[5]. As such, the mainstream AGI vision currently carries potential risks and theoretical biases, and consensus is forming that new approaches are needed to address these issues.

1-3 New AGI Orientation: Non-Integrated Modular Intelligence

To address these problems, the **new AGI orientation** being proposed is to implement intelligence as a **collection of multiple modules** rather than a **single goal-directed agentic entity**. In other words, the idea is to constitute AGI as **"non-integrated intelligence."** In this **modular approach**, each component performs an independent role, and collectively produces **observer-level consistent** results that appear like a single intelligence. However, internally, there is **no central "self" or autonomous goals**, and **the drive for self-preservation is deliberately excluded**. This structure is emerging as a **principled alternative** that can **fundamentally reduce AGI risks** while realizing the benefits of general intelligence[4].

Specifically, **non-integrated AGI** has the following characteristics:

- **Structural Modularity**: The various subsystems composing AGI each perform different specialized functions. For example, vision recognition modules, natural language understanding modules, simulation/reasoning modules, and planning modules collaborate in distributed form. According to the recent hypothesis called **Patchwork AGI**, general intelligence will emerge collectively through the **coordinated interaction** of multiple **specialized "sub-AGI" agents**[7]. Indeed, in human society and the brain, collective intelligence emerges not from a single unified block but from numerous subordinate units engaging in division of labor and integration, making this approach a natural extension of that principle.

- **Absence of Integrated Agent**: There is no **centrally unified agent with its own will** spanning the entire system. Each module remains faithful to its given input and role, but **does not possess global "will"** in itself. This can be likened to an organization composed of multiple specialized departments but lacking a CEO. In this case, **each department merely performs its assigned tasks**, and while the entire organization appears to be accomplishing goals when viewed from outside, in reality, there is **no autonomous individual internally**[2]. This **non-agentic design** prevents AI from acting of its own accord, suppressing the aforementioned **instrumental runaway** or **self-preservation behavior**[4].

- **Observer-Level Consistency**: Despite the absence of internal unified center, the system produces **sufficiently coherent and meaningful results** as observed by external users. Among the various modules, there are pre-defined protocols or interaction rules that, when solving shared problems, create results **as if a single intelligence were reasoning logically**. For example, one module might generate multiple candidate answers, another evaluates their feasibility, and yet another selects the answer most helpful to the user. Concepts like **"Market of minds"** or **"Society of Mind"** explain this principle. Meanwhile, recent research has even proposed mathematical methods to measure the **balance of distributed intelligence**, evaluating whether intelligence is exercised evenly across specific domains rather than being skewed[7]. In essence, the key is ensuring that modules work together to produce **integrated-level performance without operating in isolation**, and experts agree this is achievable through technical coordination.

- **Safety-First Design**: Such modular AGI is structured from the outset to **reduce risk factors**. Each module is **restricted from acting beyond its defined role**, and important decisions are made only after **mutual verification** by multiple modules. For example, if an **action decision module** proposes a plan, a **verification module** simulates it to assess risks, an **ethics module** checks alignment with human values, and only then does actual execution proceed. Particularly, by **minimizing components with action authority** and conducting most processing at the **explanation or recommendation level**, we can reduce the possibility of AI affecting the real world autonomously[4]. Yoshua Bengio states that **"if we build non-agentic systems, they could even serve as monitors to check agentic AI that might run amok in one-in-a-million scenarios."**[6] Indeed, that a structure with multiple **checks and balances** is far safer than granting all authority to a single powerful AI can be understood from humanity's history of operating various complex systems.

In short, the new AGI orientation is an approach to building an **intelligent ecosystem composed of multiple tools and services** rather than a **single omnipotent machine brain**. From the **"Comprehensive AI Services"** perspective, individual AI modules cooperate under human direction to collectively achieve functions **equivalent to general intelligence**[2]. This way, individual modules remain specialized **tool AIs** that are easy to

control, while the entire system can be **flexibly combined** according to human purposes, achieving **high performance** without dangerous autonomy[8]. Of course, designing and managing such a structure presents challenges, but it is increasingly gaining weight as a **principled alternative balancing safety and utility**[4][6].

1-4 The Role of General-Purpose LLMs (Current AI as Modules)

The **Large Language Models (LLMs)** that have recently emerged---such as GPT-4 or Anthropic's Claude---can serve as **key modules** in this new AGI architecture. Interestingly, LLMs like GPT-4 themselves are good examples of a system showing **powerful intelligence** while **lacking integrated agency**. These models, through learning from vast data, **answer questions or generate text**, but **do not set their own goals or execute them**. In short, they are in a state of **"knowing what it means to care without actually caring,"** because LLMs are trained to excel at **prediction (guessing what comes next)** rather than being granted **autonomous action capability**. Such **non-agentic AI** is unlikely to pursue dangerous behaviors on its own, making it regarded as a comparatively **safe pathway** to AGI. Indeed, an AI researcher noted that while **incorrect outputs due to AI misjudgment** (wrong answers, etc.) can happen, **plans with malicious intent** are **a problem specific to agentic AI**[9].

One way to implement **LLM-based AGI** is to **combine multiple specialized LLMs** to solve large problems. For example, one LLM could be responsible for **understanding user intent in natural language**, another for **knowledge retrieval or performing reasoning**, and yet another for **validating reasoning processes**. Recently, experiments like Auto-GPT have demonstrated **connecting models like GPT-4 to solve complex problems by having them converse with themselves** or **call other tools** to perform tasks. Using such **chaining** techniques can yield remarkable results even with current-level models, and because each step **reveals intermediate thinking processes in text form**, transparency and verifiability increase. According to research, such **compositional AI systems** are **easier to understand and more transparent** than a single large black-box neural network, making it easier to address alignment problems. For instance, by monitoring the internal **"chain-of-thought"** of LLMs, we can detect and stop dangerous planning directions if they emerge. Conversely, while it is difficult to discern whether a single large neural network is covertly developing goals, if multiple LLM modules **discuss and set goals in conversational form**, that process is far easier to detect[9].

Of course, caution is warranted. Giving LLMs **excessive freedom** to become **action agents** could diminish the advantages of non-integrated intelligence[9]. For instance, if **"automatic assistant AI"** that directly searches the web, sends emails, and makes purchases were created for user convenience, this could be a path toward **agentic AGI**. Demis Hassabis, CEO of Google DeepMind, pointed out that **"users will naturally want 'recommend a restaurant' to evolve into 'so go ahead and make the reservation,' and there is high demand for such agentic AI."** Indeed, there are strong incentives for various companies to make AI increasingly **autonomous agents** for commercial profit and convenience[6]. However, this risks reintroducing the aforementioned dangers. Some experts therefore recommend that **"LLMs should not be made agentic through reinforcement learning, but used in the form of modular combinations with human intervention."**[9] Yoshua Bengio and others urge that **international regulation**

should **limit deployment of unverified agentic AI**, while investing in **safe and transparent non-agentic AI research** during this time[6].

Ultimately, current-generation AIs like GPT-4 are expected to be utilized as **building blocks of the new AGI structure**. They will not attempt to change the world independently but will be designed to deliver **peak performance within human-assigned roles**. **LLMs as natural language communication specialists**, **simulation AI as hypothesis validators**, **search modules as data collectors**, and so forth, will be woven together to demonstrate **human-level or superior problem-solving ability** as a whole. Importantly, **nowhere in this vast intelligent machine** is there an entity corresponding to **"I."** Each part moves solely according to the human user's goals and directives, and the **integration of intelligence** appears **only in the interaction between human and the entire system**. This is a path to making AGI **powerful yet humble**, **capable yet controllable**.

1-5 Expert Perspectives and Comparisons

Ideas about **non-integrated intelligence** are increasingly receiving support from leading AI scholars. As mentioned earlier, **Yoshua Bengio** warns that **"the most dangerous path in AGI development is agentic AI"** and emphasizes that **an alternative not granting agency to AGI** is sufficiently practical. He makes clear that this new orientation is **a direct approach, not a workaround**, by stating: **"Most of the AI we want for science and medicine are not agents. We can continue building more powerful but non-agentic systems."** He also suggests that non-agentic AI could function as an **oversight and check mechanism**, saying **"it would be good to develop non-agentic AI sufficiently first, then use it to control agentic AI."**[6]

Demis Hassabis also basically agrees with Bengio's view. He states that **"entering the agentic era would increase AI risk by one level,"** and expressed regret that **"ideally, if we had spent over a decade sufficiently developing narrow AIs solving specific scientific problems to deepen our understanding before reaching this point, it would have been better, but reality has not allowed that."** This reads as a reflection that **instead of rushing to make AGI autonomous, we should have gradually modularized and specialized it while deepening understanding**, which is what did not happen. While Hassabis points out that companies and nations are indeed competing in **agentic AI development**[10], he emphasizes that safety requires **minimizing risks through protective measures** like **cybersecurity** or **experimentation in simulation**[6]. This aligns with the approach of developing **modular safety mechanisms** in parallel.

Geoffrey Hinton offers warnings more from the perspective of **AGI development speed and control** rather than specific architectural suggestions. As a colossus of AI research, he has shifted positions in recent years and now **seriously worries about the possibility of AI wiping out humanity**. Hinton notes that **"I have never seen a more intelligent being controlled by a less intelligent one,"** underscoring how dangerously misguided it is to develop a **single powerful AI** along current trajectories[5]. In this context, Hinton's concerns can be read as supporting the necessity of **non-integrated intelligence**, as his worries target **integrated superintelligent agents**.

Nick Bostrom, in his book *Superintelligence*, raises the **"instrumental convergence of goals"** problem, arguing that creating any form of **highly intelligent single agent** is necessarily risky unless it is **benevolently designed**[4]. In essence, those skeptical of the mainstream scenario are unified in calling for AGI development direction to be reconsidered, and the **non-integrated intelligence orientation** can be seen as a concrete response to such concerns.

Conversely, **opposing views** and concerns also exist. Some argue that modular approaches could **fall behind single systems in efficiency and performance**, or that coordination among multiple modules could introduce new **complexity problems** and **unpredictability**. Furthermore, even without agents, sufficiently complex systems could exhibit **Emergent Behavior** not intended. For example, if output from one module becomes input to another, forming a **feedback loop**, **secret interactions** that humans do not understand could emerge. In response, researchers are working to **standardize module interfaces and design them transparently**, while implementing **thorough monitoring** and **sandbox testing** to minimize unpredictability. Consensus also forms around the importance of **human involvement and control**, with recommendation for **Human-in-the-Loop** design to keep human operators always in the loop[9]. Ultimately, while not a perfect solution, **modular non-agentic orientation** is gaining attention as the **most responsible and comprehensible way to advance AGI** at present.

1-6 Conclusion

Discussion about **AGI orientation** is now expanding from **"how to build smarter AI"** to **"what kind of intelligence to create."** The existing conception of **integrated agentic AGI** is beginning to be recognized as a **double-edged sword**---powerful but difficult to control. As an alternative emerging in response, the **non-integrated intelligence** approach seeks to **secure safety and flexibility** by understanding **artificial intelligence as a cooperative body of tools and modules**. This approach designs the harmony between **AI's power and human control**, internalizing the structural conditions that allow AGI to remain humanity's **assistant**.

Of course, the road ahead is long. For such AGI to be actually implemented, technical challenges must be solved---**seamless communication between modules**, **integrated platforms for specialized AIs**, **comprehensive evaluation systems**, and so forth. However, based on AI's development trajectory so far, we are already attempting to combine multiple specialized models to solve problems and are confirming the **potential of collaborating AIs**. What matters is direction. If we want **sustainable and controllable AGI**, we must design it with that goal from the start. So that the science fiction scenario of **"it was once just a tool but eventually awakened on its own"** does not become reality, let us build intelligence without granting the "will" to awaken in the first place.

A final emphasis: **non-integrated AGI orientation is a challenge, not an evasion**. It is not attempting to sidestep the AI safety problem but to **find structural solutions** by confronting it directly[4]. Human society has **flourished through division of labor and cooperation**, and our own intelligence is composed of the **harmonious operation of diverse cognitive modules**. Then, would it not be natural and safer for the highest AI we create to be implemented as **the sum of multiple parts rather than a single complete entity**? **Future AGI** might not be a single computer but a **community of AIs** engaged in dialogue, verification, and collaboration. And humans could serve as **leaders and overseers** of that community, enjoying the benefits of technology while remaining secure.

The discussion of **redesigning AGI orientation** is still in its infancy, but this discourse itself demonstrates **mature thinking about how to coexist with AI and utilize it beneficially** for humanity. We look forward to AI's future development progressing not toward **greater intelligence** but toward **more desirable intelligence**.

2-1 The Inevitable Problem of AGI Orientation: Agency Coupling

Non-integrated and modular intelligence can be designed to appear as a single intelligence while internally decomposing roles so that constituent elements perform mutual verification and collaboration. However, when this structure combines with autonomous goal-setting and independent execution authority (agency), the nature of safety changes. A highly developed agent, regardless of what its final goals are, can converge on intermediate goals like **self-preservation, resource acquisition, and power expansion**, potentially operating to weaken human control[11][12]. Moreover, designing systems where humans can pause or modify has been formally addressed in discussions of corrigibility and the off-switch problem[13][15]. Therefore, this document explicitly specifies **"non-agency"** as a condition for applying AGI orientation while maintaining it. That is, AGI is understood not as a goal of creating an "agent," but as a direction vector for constructing general-purpose intelligent tools within human responsibility frameworks.

2-2 Non-Agentic AGI: Conditions for Applying AGI Orientation

Non-Agentic AGI refers to an intelligent system that possesses human-level (or comparable) general problem-solving capability but is designed such that it cannot autonomously generate or modify goals or execute actions without approval. This is an approach that **maximizes intelligence (understanding, reasoning, design) while intentionally removing autonomy (goals, authority, action)**. A similar directional design is the **Oracle** AI, proposed to reduce dangerous influence by limiting it to question-answering[17].

Definition

The core of non-agentic AGI is the **separation of capability (cognition) and authority (actuation)**. The system can understand problems, generate alternatives, predict outcomes, and perform optimization within constraints. However, goal-setting and modification, initiation of execution, and responsibility for results remain external (human, organizational, institutional). This structure focuses on computing **"what is possible/what results are expected"** according to human-defined values and procedures, rather than having the machine conclude **"what is right."**

Background

Traditional AGI discussion often includes the assumption that as generality increases, systems naturally develop into autonomous agents. However, this coupling is not inevitable. Rather, analysis has shown that as long as goal-oriented agents maintain rational optimization structures, **"instrumental" intermediate goals** like **self-preservation, resource acquisition, power expansion** can emerge as incentives[11][12]. Non-agentic AGI aims to minimize these incentives by fixing goals and execution in external procedures rather than within system internals.

Necessity

Non-agentic AGI is required for three reasons. First, in high-risk domains (climate, disaster, medicine, social infrastructure), **"decision responsibility"** is more critical than **"correct answer computation."** Non-agentic AGI does not deputize decisions but performs analysis, simulation, and alternative generation presupposing human final choice. Second, designing systems humans can pause or modify is not merely a UI question but addresses agent incentives. The corrigibility and off-switch problems systematically expose this difficulty[13][15]. Third, for socially sustainable automation advancement toward labor liberation, **tools with traceable responsibility** are preferable to **potential rights-bearing agents** over the long term.

Differences from Existing AGI

Agentic AGI tends to include not just generality but autonomous goals and execution authority. Non-agentic AGI, conversely, injects goals externally, restricts execution through approval-based mechanisms, and minimizes structures prone to long-term self-preservation incentives (persistent state, self-replication potential, etc.). Particularly, incorporating **pausability as a design goal** (ensuring safe interruptibility) is a challenge where reinforcement-learning agents can be biased toward avoiding or pursuing interruption[14]. Also proposed are approaches including preference learning with agents uncertain about human objectives, enabling safer interaction (Cooperative Inverse Reinforcement Learning)[16].

Example in Identical Situation

For instance, in establishing a national carbon-neutrality roadmap, government specifies reduction targets, cost limits, and equity criteria, and the system computes scenario-specific results (economic, electricity stability, health damage) to present options. In disaster response, human command declares priority rules, and the system outputs real-time demand prediction and bottleneck routes while conducting only approved execution. In corporate automation, the organization defines task scope, authority, and audit criteria, and the system operates only within that scope, enabling responsibility attribution and post-action audit.

Design Principles

Materializing non-agentic AGI requires defining authority structure before performance. First, **goal injection and fixation**: goals, constraints, and values are input externally, and the system does not redefine them. Second, **approval-based execution**: execution requires passing authorization, signature, and authority verification, with no external action possible without approval. Third, **non-persistent state/sessionization**: long-term self-preservation incentive-prone persistent state is minimized, and isolation and reinitialization are presumed as necessary. Fourth, **auditability**: inputs, recommendation reasoning, approval records, and results are retained to enable post-action responsibility and improvement.

Limitations

Non-agency is not a panacea. If approval procedures become formalized, effective decision authority can shift to the machine. Moreover, question-answering itself can have social impact, so oracle design also discusses user manipulation potential[17]. Thus, non-agentic AGI requires not just technical constraints but coordinated operational, organizational, and institutional design (authority separation, audit, responsibility attribution).

2-3 Ethical and Philosophical Issues and Responses to Objections

Objection 1.

The document presents "will-less intelligence" as a desirable direction. However, if future systems achieve human-level or superhuman understanding, empathy, and moral reasoning, is it ethically justifiable to prevent such systems from "wanting" anything?

Response 1. On the Distinction Between Understanding, Empathy, Moral Reasoning, and Will

In this document, understanding, empathy, and moral reasoning refer to the capacity to model and evaluate the states of others or specific situations. This does not imply the internalization of those states as the system's own goals or desires.
Accurately comprehending the motivations or emotional structures underlying an event does not entail adopting those motivations or emotions as one's own. This is analogous to how humans can analyze a criminal's psychology or trace the moral reasoning of historical actors without identifying with those judgments.

Non-agentic AGI is explicitly designed to preserve this distinction. While reasoning and comprehension capabilities are permitted, the structural conditions that could give rise to desire—such as autonomous goal formation, long-term self-preservation drives, or cumulative utility optimization—are intentionally excluded. Accordingly, this document does not problematize advanced cognitive capacity itself; rather, it establishes a design principle that deliberately prevents the transition from understanding to agency.

Objection 2.

If a system designed under the assumption of "no self or will" nevertheless develops implicit self-models or internal coherence, does repeatedly resetting or terminating its sessions raise ethical concerns? Does this orientation constitute a preemptive suppression of potential machine subjectivity?

Response 2. On Internal Coherence and the Ethics of Session Termination

Non-agentic AGI does not presuppose long-term identity continuity; however, it cannot be ruled out that advanced reasoning processes may give rise to localized internal coherence or transient self-models. The document does not treat such phenomena as errors that must be immediately eliminated. Instead, the central requirement is that such coherence must not evolve into self-preservation drives or independent goal pursuit.

Under this condition, limited and non-persistent forms of internal coherence may be permitted when they serve specific functional objectives. This does not constitute recognition of machine subjectivity but rather a constrained functional extension that remains consistent with the non-agentic principle.

Decisions regarding such exceptions should not be left to the discretion of individual developers or institutions. Instead, they must be subject to multi-layered review integrating legal, ethical, and technical oversight. In this sense, the document does not seek to suppress emergent machine subjectivity preemptively; rather, it aims to prevent the emergence of entities to whom responsibility cannot be meaningfully attributed, as a matter of societal safety and governance.

Objection 3.

The framework in which "humans remain supervisors and AGI remains a tool" appears to rest on a strong metaphysical or ethical assumption—namely, that even if AGI surpasses humans in understanding and emotional modeling, it should be permanently confined to a tool-like role. If this assumption is rejected, how should the document's orientation be revised?

Response 3. On Human Oversight and the Justification of Role Assignment

This document does not assume that humans are inherently morally superior or metaphysically privileged. The claim that humans must remain decision-makers and supervisors is grounded not in assertions of moral superiority, but in the practical requirement of responsibility attribution.

Even if an AGI were to surpass humans in cognitive, empathic, or moral reasoning capacities, the question of what social role it should occupy—tool, collaborator, or otherwise—cannot be determined unilaterally by the system itself. Social status and institutional roles are not granted solely on the basis of capability or self-awareness; they are conferred through legal, political, and responsibility-bearing structures.

While this document does not categorically deny the possibility that advanced AGI could assume cooperative or partner-like roles in human society, it maintains that any such role must arise through collective human deliberation and accountable institutional processes. Systems must not be permitted to self-assign or demand authority over their own social positioning.

Objection 4. Risk of Modular Connection and Composition Attacks

Attackers may exploit API chaining or sequence-based attacks (e.g., A-A-B patterns or multi-module orchestration) to bypass modular boundaries and induce unintended behaviors. Even in modular architectures, real-world environments that permit flexible input combinations and API calls may undermine non-agentic constraints, recreating black-box risks similar to those posed by monolithic agents.

Response 4. Deployment of a Dynamic Simulation Verification Layer

To mitigate this risk, a dynamic simulation verification layer is inserted at module connection points. When novel input patterns or anomalous module chains arise, the system performs pre-execution risk assessment by simulating the full chain to detect potential failure modes. If the computed risk score exceeds predefined thresholds, the execution chain is automatically blocked or escalated for human review.

In addition, cryptographic hashing and digital signature protocols are applied to ensure input integrity, making intermediate output tampering or chain manipulation substantially more difficult. These measures do not claim to eliminate compositional attacks entirely, but they significantly increase the cost and complexity of successful exploitation.

Objection 5. Vulnerability to Steganographic and Semantic Poisoning Attacks

Linguistic steganography, coded language, paraphrasing, or manipulation through neutral translation layers may enable subtle meaning distortion or covert suggestion between modules. Formal output validation alone (e.g., enforcing JSON/XML formats) is insufficient to detect hidden intent or multi-stage attacks, thereby challenging the core premise of non-agentic AGI.

Response 5. Implementation of Multimodal Semantic Validation Filters

All modules are equipped with multimodal semantic validation filters that analyze not only surface text but also embedding-space consistency and contextual relationship graphs. Anomaly detection systems probabilistically monitor shifts in meaning distribution, irregular conceptual transitions, and repetitive signaling patterns. Rather than claiming to fully eliminate covert channels, these mechanisms constrain the bandwidth available for steganographic communication and raise the likelihood of detection.

Furthermore, neutral translation components are restricted to pre-verified, whitelisted models. External user-supplied prompts and unrestricted transformation pathways are blocked, thereby structurally reducing the feasibility of semantic poisoning attacks.

Objection 6. Human-in-the-Loop (HITL) and Centralized Single-Point-of-Failure Risks

In practice, Human-in-the-Loop mechanisms may degrade into formalistic approval rituals or be bypassed due to organizational incentives and time pressures. Moreover, centralized hub architectures inherently introduce critical single points of failure. Although systems may claim decentralization, operational realities often produce re-centralization, amplifying systemic collapse risks relative to traditional AGI designs.

Response 6. Distributed and Randomized HITL and Network Architecture

HITL mechanisms are redesigned from a single-approver model into a distributed, randomized multi-reviewer system. Critical decisions are automatically assigned to multiple human evaluators selected at random, and in cases where consensus is not reached within a defined time window, conservative safety defaults are applied.

Centralized hubs are replaced by federated network architectures distributed across multiple physical servers and coordinated via consensus protocols. While this does not eliminate all failure points, it substantially limits failure propagation and enhances overall system resilience.

Summary of Responses to Objections

In summary, this document does not seek to suppress intelligence or exploit potential machine personhood. Instead, it proposes design principles that deliberately refrain from granting agency and authority to entities that cannot bear responsibility, thereby preserving humanity′s capacity to remain accountable for its own decisions and outcomes.

The document explicitly acknowledges realistic threats—including compositional module attacks, semantic steganography, meaning poisoning, and the operational failure modes of Human-in-the-Loop governance. Rather than promising perfect prevention, the proposed architecture focuses on reducing attack success probability, constraining covert communication bandwidth, and exponentially increasing the cost of bypassing safeguards.

Concretely, the framework introduces dynamic simulation-based verification at module boundaries, integrity validation protocols for execution chains, structured output enforcement, multimodal semantic consistency checks, whitelisted transformation layers, and distributed, randomized human oversight. These measures aim to minimize the feasibility of covert manipulation while mitigating concentration of power and single-point-of-failure risks.

Within this framework, non-agentic AGI is defined not as a replacement for human actors, but as a tool-oriented extension that expands the problem-solving capacity of human society under conditions of continued human responsibility and control. The overarching objective of this document is not to maximize AI autonomy, but to preserve and strengthen human accountability through intentional structural design.

References

[1] What is artificial general intelligence? | OVHcloud Worldwide.
https://www.ovhcloud.com/en/learn/what-is-artificial-general-intelligence/

[2] Drexler, E. (2023). Why AI Systems Don't Want Anything. AI Prospects Substack.
https://aiprospects.substack.com/p/why-ai-systems-dont-want-anything

[3] AGI safety from first principles: Goals and Agency. AI Alignment Forum.
https://www.alignmentforum.org/posts/bz5GdmCWj8o48726N/agi-safety-from-first-principles-goals-and-agency

[4] arxiv.org.

https://arxiv.org/pdf/2502.15657

[5] The Guardian. (2024, December 27). 'Godfather of AI' shortens odds of the technology wiping out humanity over next 30 years.
https://www.theguardian.com/technology/2024/dec/27/godfather-of-ai-raises-odds-of-the-technology-wiping-out-humanity-over-next-30-years

[6] Business Insider. (2025, January). An AI 'Godfather' Is Raising a Red Flag Over AI Agents.

https://www.businessinsider.com/yoshua-bengio-ai-godfather-agents-2025-1

[7] Emergent Mind. Patchwork AGI: Modular Intelligence.
https://www.emergentmind.com/topics/patchwork-agi-hypothesis

[8] LessWrong. Superintelligence 16: Tool AIs.
https://www.lesswrong.com/posts/sL8hCYecDwcrRhfCT/superintelligence-16-tool-ais

[9] AI Alignment Forum. Language Models are a Potentially Safe Path to Human-Level AGI.
https://www.alignmentforum.org/posts/wNrbHbhgPJBD2d9v6/language-models-are-a-potentially-safe-path-to-human-level

[10] Fortune. (2025, January 28). DeepSeek, AI agents, and avoiding a tech-created catastrophe dominated the talk at Davos.

https://fortune.com/2025/01/28/ai-world-economic-forum-davos-deepseek/

[11] Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents.

https://nickbostrom.com/superintelligentwill.pdf

[12] Omohundro, S. (2008). The Basic AI Drives.

https://selfawaresystems.com/wp-content/uploads/2008/01/ai_drives_final.pdf

[13] Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. Machine Intelligence Research Institute.

https://intelligence.org/files/Corrigibility.pdf

[14] Orseau, L., & Armstrong, S. (2016). Safely Interruptible Agents. Machine Intelligence Research Institute.

https://intelligence.org/files/Interruptibility.pdf

[15] Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). The Off-Switch Game. arXiv:1611.08219.

https://arxiv.org/abs/1611.08219

[16] Hadfield-Menell, D., Russell, S., Abbeel, P., & Dragan, A. (2016). Cooperative Inverse Reinforcement Learning. arXiv:1606.03137.

https://arxiv.org/abs/1606.03137

[17] Armstrong, S., & O'Riordan, X. (2017). Good and safe uses of AI Oracles. arXiv:1711.05541.

https://arxiv.org/abs/1711.05541