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

English

Table of Contents

1-1 Introduction

Today, artificial general intelligence (AGI) is commonly imagined as a single, unified artificial intelligence with human-level intelligence [1]. Many people picture a solitary agent that moves with autonomous goals, like the AI characters seen in films. Because human and animal intelligence, born of biological evolution, includes a sense of self, instinctive goals, and a drive for self-preservation, we intuitively expect powerful AI to likewise pursue its own goals and protect itself as a single agent [2]. However, this intuitive expectation has limitations and risks. This document examines the limitations and risks of traditional AGI concepts and proposes a new AGI orientation: a non-integrated, modular intelligence. This approach excludes a unified agency and aims for a structure that elicits coherent intelligent behavior from an observer perspective through the combination of multiple specialized roles. It is presented not as a stopgap but as a safer and more principled alternative to agentic AGI.

1-2 Conventional AGI Concepts and Their Limitations

AGI is generally defined as "an AI that can perform all intellectual tasks that humans can" [1]. This definition typically carries an implicit assumption: a single autonomous agent in which multiple capabilities are integrated. For example, classical notions of intelligence include self-awareness,

long-term planning, outcome prediction, and an internally unified goal (single agency) [3]. However, this integrated-agency view involves two major problems.

First, there is the problem of safety and control. According to research by philosophers such as Nick Bostrom, sufficiently intelligent agentic AI is likely to pursue instrumental goals such as self-preservation and resource acquisition regardless of its final goal [4]. This is known as instrumental rationality convergence and raises concerns that an AI may prevent itself from being shut down or seek more resources to achieve its goal [5]. Bostrom warns that a powerful AGI could become a new intelligent "species" beyond human control and pose an existential threat to humanity [6]. Even if its initial intent is harmless, an AGI with an internal goal system may, over time, reinterpret goals in unexpected ways or pursue its own decision-making [7]. The worst-case scenarios in AI safety discussions are largely connected to situations in which such autonomous agentic AGI conflicts with humans or escapes human control [8].

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 "biological intelligence made self-preservation a necessary condition through evolution, but modern AI does not" [2][9]. In other words, current AI systems do not have an instinct to preserve themselves, nor do they need one; they are developed based on task performance defined by humans [10]. Nevertheless, expecting autonomy and self-preservation in AGI by using human intelligence as the reference may be a misplaced biological bias. For example, powerful AI systems today, such as GPT-4 and AlphaFold, demonstrate very high intellectual performance without their own desires or survival instincts. Therefore, the dominant definition "strong AI = integrated agent" is not a necessity but a choice, and it may be a choice that increases risk.

For these reasons, major scholars in the field have also raised warnings. Deep learning pioneer Yoshua Bengio notes that "all catastrophic scenarios of AI occur when there is an agent, and it is possible to reach AGI without agency" [11]. Bengio calls agentic AI "the most dangerous path" and emphasizes that beneficial work in science and medicine can be achieved without granting autonomy to AI [12]. In the same context, Geoffrey Hinton has expressed concern that "the probability that AI will wipe out humanity within the next few decades is 10-20%" and reminded us that history provides no example of a less intelligent being controlling a more intelligent one [13][14]. In this way, the mainstream vision of AGI carries potential risks and theoretical bias, and a consensus is forming that a new approach is needed.

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

To address the problems above, the proposed new AGI orientation is to realize intelligence not as a single goal-directed agent but as a collection of multiple modules. In other words, it proposes constructing AGI as "non-integrated intelligence." In this modular approach, each component performs an independent role, and as a whole it produces coherent results at the observer level as if it were a single intelligence. Internally, however, there is no central "self" or intrinsic goal, and the drivers of self-preservation are intentionally excluded. This structure is emerging as a principled alternative that can realize the benefits of general intelligence while fundamentally reducing AGI risk [15][16].

Specifically, non-integrated AGI has the following characteristics:

Structural modularity: Multiple subsystems that constitute AGI each take on specialized functions. For example, visual perception modules, natural-language understanding modules, simulation/reasoning modules, and planning modules can collaborate in a distributed form [17]. According to a recent hypothesis called Patchwork AGI, general intelligence may emerge collectively through coordinated interaction among multiple specialized "sub-AGI" agents [17]. In fact, human society and the brain are also not a single block but manifestations of intelligence

through repeated division of labor and integration among countless subunits, making this approach a natural extension.

Absence of an integrated agent: There is no central agent, unified across the whole system, that possesses its own will. Each module is faithful to its given inputs and role, but by itself does not have a global "will" [10]. This can be likened to an organization composed of specialized departments but lacking a CEO: each department performs its assigned task, and from the outside the organization appears to achieve its objective, but internally there is no autonomous individual [10]. Such a non-agentic design eliminates room for the AI to act on its own, suppressing instrumental runaway or self-preservation behavior mentioned earlier [16].

Observer-level coherence: Even without an integrated center, the system can produce sufficiently coherent and meaningful results from the perspective of external users. Predefined protocols and interaction rules among modules allow the system to generate outputs that look like a single intelligence reasoning logically when solving a shared problem [18]. For example, one module may generate multiple candidate answers, another may evaluate plausibility, and another may select the response most helpful to the user. Concepts such as a "Market of minds" or a "Society of Mind" explain this principle. Recent work even proposes mathematical methods to measure balance in distributed intelligence so that it is not overly skewed toward a particular domain [19][20]. In short, the key is to ensure that modules do not operate in isolation but collectively achieve near-integrated performance, which experts argue is feasible through technical coordination.

Safety-first design: Such modular AGI can be structured to reduce risk factors from the outset. Each module is constrained from acting outside its assigned role, and important decisions are made through mutual verification among multiple modules. For example, if an action-selection module proposes a plan, a verification module can simulate and assess risk, and an ethics module can check alignment with human values before any execution occurs. In particular, minimizing the components with actuation authority and making most processing remain at the level of explanation or recommendation can reduce the likelihood that AI affects the real world on its own [16]. Yoshua Bengio has noted that "if we build non-agentic systems, they could act as monitors to keep in check agentic AI that might run amok" [21]. The lesson that distributing authority with checks and balances is safer than concentrating it in a single powerful entity is also supported by the history of operating complex systems in human society.

In sum, the new AGI orientation is an approach that builds an ecosystem of intelligence consisting of multiple tools and services rather than creating a single omnipotent machine brain [10]. Under the "Comprehensive AI Services" view, individual AI modules cooperate under human direction to collectively achieve functionality comparable to general intelligence [10]. In this way, each module remains a specialized tool AI that is easier to control, and the overall system can be flexibly assembled according to human purposes, enabling high performance without dangerous autonomy [22][23]. While designing and managing such a structure is challenging, it is increasingly seen as a principled alternative that seeks to secure both safety and utility [24][11].

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

Large language models (LLMs) that have emerged today - such as GPT-4 and Anthropic's Claude - can serve as core modules in this new AGI architecture. Notably, LLMs such as GPT-4 are good examples of systems that exhibit very strong intelligence while lacking integrated agency. These models answer questions and generate text through large-scale training, but they do not set goals or execute actions on their own. In short, they do not "care" but can "know what it means to care" [25]. This is because LLMs are trained to predict the next token, not to act with autonomy [26]. Such non-agentic AI is therefore considered a relatively safe path toward AGI, as it is unlikely to attempt dangerous actions by itself [27]. As one AI researcher has pointed out, incorrect outputs may occur due to AI error, but malicious intentions and planning are primarily a concern for agentic AI [27].

One way to implement LLM-based AGI is to combine multiple specialized LLM modules to solve large problems. For example, one LLM can interpret user intent in natural language, another can retrieve knowledge or perform reasoning, and another can verify the reasoning process. Recently, experiments have also been conducted to chain models like GPT-4 to converse with themselves to solve complex problems or to call other tools to perform tasks (e.g., Auto-GPT). Such chaining techniques can produce remarkable results even with current models [28], and because each step exposes intermediate reasoning in text form, transparency and verifiability can be improved [29][27]. Research suggests that such compositional AI systems are easier to understand and more transparent than a single large black-box neural network, making alignment (goal alignment) easier to address [30][31]. For example, monitoring LLMs' internal chain-of-thought can allow us to detect and halt planning if it moves in a dangerous direction [27]. In contrast, while it is difficult to detect whether a single large neural network is secretly forming goals, it is much easier to observe such processes when multiple LLM modules discuss and formulate goals in a conversational structure [27].

Of course, there are caveats. If we grant LLMs excessive freedom and turn them into action agents, the advantages of non-integrated intelligence can be weakened [32]. For instance, if an "AI assistant" is made to directly browse the web, send emails, and purchase items for user convenience, this may gradually become a path toward agentic AGI. Demis Hassabis, CEO of Google DeepMind, has pointed out that users naturally move from "recommend a restaurant" to "then book it for me, too," and that demand for agentic AI is high [33]. In reality, commercial incentives and convenience create strong pressure for companies to make AI increasingly autonomous agents [34]. However, this can reintroduce the risks discussed above, so some experts recommend that "we should not turn LLMs into agents via reinforcement learning; instead, we should use them in a modular composition with human involvement" [35]. Bengio and others call for international regulation to limit deployment of unverified agentic AI and to focus meanwhile on research into safe and transparent non-agentic AI [36].

Ultimately, contemporary AI such as GPT-4 is expected to serve as building blocks for a new AGI structure. These systems will not try to change the world on their own, but will be designed to achieve maximum performance within roles assigned by humans. LLMs as language communication specialists, simulation AIs as hypothesis verifiers, and search modules as data collectors can be combined so that the system as a whole demonstrates human-level or higher problem-solving capability. The critical point is that nowhere in this large machine intelligence is there an entity corresponding to "I." Each part moves only according to the goals and instructions of human users, and the integration of intelligence appears only in the interaction between humans and the overall system. This can be seen as a path to making AGI powerful yet humble, capable yet controllable.

1-5 Expert Perspectives and Comparative Views

Ideas about non-integrated intelligence are gaining support from an increasing number of AI scholars. As noted earlier, Yoshua Bengio warns that "the most dangerous path in AGI development is agentification" and emphasizes that alternatives without agency are sufficiently practical [11][12]. By stating that "most of the AI we want for science and medicine is not agentic; we can keep building more powerful yet non-agentic systems" [11], he makes clear that this orientation is not a detour but a direct path. He also suggests that "it is good to develop non-agentic AI sufficiently first and use it to control agentic AI" [21], implying that modular AI can function as a supervisory check-and-balance mechanism.

Demis Hassabis also largely agrees with Bengio's view. He stated that "the moment we enter the agentic era, AI risk increases by a level" [37], and expressed regret that, ideally, "it would have been good if we had advanced narrow AIs solving specific scientific problems for more than 10

years to buy time to deepen our understanding, but reality did not turn out that way" [38]. This can be read as a reflection that, rather than rushing to make AGI autonomous, we should have proceeded step by step with modularization and specialization to increase understanding. Hassabis nevertheless points out that companies and countries are in an agentic AI race [39], while emphasizing that we must minimize risk through safeguards such as cybersecurity protections and experiments within simulations [34]. This aligns with approaches that develop modular safety mechanisms in parallel.

Geoffrey Hinton offers warnings less about specific architectures than about the speed of AGI development and the difficulty of control. As a leading figure in AI research, he has shifted his stance in recent years and now "seriously worries about AI causing human extinction" [13]. Hinton has remarked that "we have never seen an example of a less intelligent being controlling a more intelligent one" [40], underscoring how dangerous it could be to continue toward a single powerful AI along current trajectories. In this context, Hinton's concern targets integrated superintelligent agents, and thus can be seen as indirectly supporting the need for non-integrated intelligence. Nick Bostrom likewise raises the issue of "instrumental goal convergence" in Superintelligence, arguing that creating a highly intelligent single agent in any form is inherently risky unless it is designed to be aligned [5]. In short, skeptics of mainstream scenarios converge on the call to "reconsider the direction of AGI development," and the non-integrated intelligence orientation can be viewed as one concrete answer to such concerns.

Meanwhile, there are also opposing views and concerns. Some argue that modular approaches may lag behind single systems in efficiency and performance, or that coordinating multiple modules may introduce new complexity and unpredictability. Even without an agent, sufficiently complex systems may still exhibit unintended emergent behavior. For example, if one module's output becomes another module's input and forms feedback loops, there may be opaque interactions that humans do not understand. To address this, researchers seek to standardize and make module interfaces transparent, and to minimize unpredictability through rigorous monitoring and sandbox testing [27]. There is also broad agreement that human involvement and control remain important, recommending human-in-the-loop designs in which a human operator stays within the loop [35]. While not a perfect solution, the modular non-agentic orientation is drawing attention as the most responsible and understandable approach to advancing AGI under current conditions.

1-6 Conclusion

Discussion of AGI orientation is expanding beyond "how to build smarter AI" to "what form of intelligence to build." The conventional integrated-agent vision of AGI is powerful but is increasingly recognized as a double-edged sword that is hard to control. As an alternative, the rising non-integrated intelligence approach seeks to secure safety and flexibility by treating AI as a cooperative system of tools and modules. This approach designs for harmony between the power of intelligence and human control, embedding structural conditions for AGI to remain a helper to humanity.

Of course, there is still a long way to go. For such AGI to be implemented in practice, technical challenges such as smooth communication among modules, an integration platform for specialized AIs, and comprehensive evaluation systems must be solved. However, recent trends in AI development show that we are already attempting to solve problems by combining multiple specialized models and are confirming the potential of collaborative AIs. The key is direction. If we want sustainable and controllable AGI, we must design it with that goal from the beginning. To prevent the science-fiction plot of "it started as a tool but eventually awakened on its own" from becoming reality, we must build intelligence that does not grant the very "will" that could awaken.

Finally, it should be emphasized that non-integrated AGI orientation is not avoidance but a challenge. It does not seek to sidestep AGI safety problems, but to confront them directly and find

structural solutions [24]. Human society has prospered through division of labor and cooperation, and our own intelligence is composed of the harmonious operation of diverse cognitive modules. If so, would it not be more natural and safer for the best AI we create to be realized as a sum of parts rather than a single complete entity? The AGI of the future may be not a single computer but a community of AIs that talk, verify, and work together. Humans, as the leader and supervisor of that community, could enjoy the benefits of technology while remaining confident in safety. The discussion of redesigning AGI orientation is still in its early stages, but the emergence of such discourse is a positive sign of a mature concern for how to make AI coexist with humanity and be used for good. We can hope that future AI development moves not toward greater intelligence, but toward more desirable intelligence.

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

Non-integrated, modular intelligence can be designed so that it appears as a single intelligence externally while internally decomposed components perform mutual verification and collaboration. However, when this structure is coupled with autonomous goal-setting and independent execution authority (agency), the nature of safety changes. Advanced agents can converge on intermediate goals such as self-preservation, resource acquisition, and power expansion regardless of their final goals, and these tendencies can weaken human control [51][52]. Moreover, the difficulty of designing systems that allow human intervention (shutdown and modification) has been formalized and discussed through corrigibility and the off-switch problem [53][55].

Accordingly, this document maintains "AGI orientation (securing general problem-solving capability)" while explicitly stating "non-agency (non-agentic)" as a condition for applying it in the real world. In other words, AGI is understood not as the goal of creating an "agent," but as a direction vector for composing general intelligence tools that can be used within a human responsibility framework.

2-2 Non-Agentic AGI: A Condition for Applying AGI Orientation

Non-Agentic AGI refers to an intelligence system that has human-level (or comparable) general problem-solving capability but is designed so that it cannot generate or change goals on its own, nor execute without approval. This is an approach that maximizes "intelligence (understanding, reasoning, design)" while intentionally removing "autonomy (goals, authority, action)." A related proposal is Oracle AI, which limits the system to "question-and-answer" to reduce dangerous influence [57].

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 under constraints. However, the setting and modification of goals, the initiation of execution, and responsibility for outcomes remain external (humans, organizations, institutions). Rather than having the machine conclude "what is right," this structure focuses on computing "what is possible / what outcomes are expected" under values and procedures defined by humans.

Background

Traditional AGI discussions often include the assumption that as generality increases, systems naturally develop into autonomous agents. However, this coupling is not inevitable. Rather, analyses suggest that as long as goal-directed agents have a rational optimization structure, incentives for instrumental intermediate goals such as self-preservation, resource acquisition, and power expansion can arise [51][52]. Non-Agentic AGI aims to minimize these incentive structures by fixing goals and execution not inside the system but through external procedures.

Necessity

The need for Non-Agentic AGI can be summarized in three points. First, in high-risk domains (climate, disasters, healthcare, social infrastructure), "responsibility for decisions" matters more than "computing answers." Non-Agentic AGI does not make decisions on behalf of humans; it performs analysis, simulation, and option generation under the premise that humans make the final choice. Second, designing systems that humans can stop or modify is not merely a UI problem but a problem of agent incentives. Corrigibility and the off-switch problem systematically reveal this difficulty [53][55]. Third, to pursue labor liberation (automation) in a socially sustainable way, tools with clear responsibility attribution are preferable to agents that could become rights-bearing subjects over time.

Differences from Conventional AGI

Agentic AGI tends to include not only generality but also autonomous goals and execution authority. By contrast, Non-Agentic AGI injects goals externally, limits execution via approval, and minimizes structures that easily create incentives for long-term self-preservation (persistent state, potential for self-replication, etc.). In particular, approaches that include "interruptibility" as an explicit design goal (e.g., safe interruptibility) have been studied as methods to prevent reinforcement-learning agents from avoiding or seeking interruptions [54]. In addition, Cooperative Inverse Reinforcement Learning (CIRL) has been proposed as an attempt to induce safer interaction by incorporating preference learning into agents that are uncertain about objectives [56].

Examples in Identical Contexts

For example, when developing a national carbon-neutral roadmap, the government specifies emissions targets, cost caps, and equity criteria, and the system computes scenario outcomes (economy, grid stability, health impacts) to present options. In disaster response, human leadership declares priority rules, and the system produces real-time demand forecasts and bottleneck routes, while execution proceeds only for approved plans. In enterprise automation, the organization defines the scope of work, authority, and audit standards, and the system performs tasks only within that scope, enabling responsibility attribution and post-hoc auditing.

Design Principles

To realize Non-Agentic AGI, the authority structure must be defined before performance considerations. First, goal injection and fixation: goals, constraints, and value criteria are provided

externally, and the system does not redefine them. Second, approval-gated execution: execution must pass approval, signatures, and authorization checks, and unapproved external intervention is disallowed. Third, non-persistent state / sessionization: persistent state that tends to create long-term self-preservation incentives is minimized, and isolation and reinitialization are assumed when needed. Fourth, auditability: inputs, reasoning for recommendations, approval records, and outcomes are logged to enable accountability and improvement.

Limitations

Non-agency is not a panacea. If approval procedures become formalities, de facto decision authority can migrate to the machine. Moreover, even "question-and-answer" can carry social influence, and Oracle designs discuss the possibility of manipulating users [57]. Therefore, Non-Agentic AGI requires not only technical constraints but also operational, organizational, and institutional design (separation of authority, auditing, responsibility attribution).

2-3 Ethical and Philosophical Objections and Responses

Objection 1. The document presents "intelligence without will" as an ideal direction, but in the long run, can it be ethically justified to build systems that perform fine-grained understanding, empathy, and moral reasoning at or above the human level while making them unable to "want" anything?

Objection 2. If a system designed on the premise of "no self / no will" in fact forms an implicit self-model or internal coherence, does the document provide any criteria for whether it is morally permissible to repeatedly reset and end sessions in that state? Is this orientation not a preemptive suppression of possible machine subjectivity?

Objection 3. The stance that "humans are the leaders and supervisors, and AGI is forever a tool" seems to require a strong metaphysical/ethical premise: that even if we could build entities that surpass humans in understanding and emotional modeling, we would treat them only as tools forever. If this premise is rejected, how should the document's overall orientation be revised?

Response to Objection 1. On distinguishing understanding/empathy/moral reasoning from will

In this document, understanding, empathy, and moral reasoning refer to the capacity to model and evaluate a situation or another party's state, and this does not immediately imply internalizing that state as one's own purpose or desire. Simply understanding the motives and emotional structure of an event in detail does not mean that those motives and emotions become one's own desires. This is analogous to how humans can analyze a criminal's psychology or trace a historical actor's moral judgment without identifying with it.

Non-Agentic AGI is designed so that this distinction is maintained. In other words, understanding and reasoning are permitted, but the conditions under which desire could arise - autonomous goal generation, long-term self-preservation, and cumulative utility optimization - are structurally removed. Accordingly, this document does not problematize advanced understanding capability itself; it adopts, as a design principle, the intentional blocking of the point at which understanding transitions into agency.

Response to Objection 2. On internal coherence formation and the ethics of session termination

Non-Agentic AGI does not assume long-term continuity of self, but it cannot be ruled out that localized internal coherence or a temporary self-model may form during advanced reasoning. The document does not treat such phenomena as errors that must be immediately eliminated. However, the key condition is to prevent such coherence from transitioning into self-preservation or independent goal pursuit.

Accordingly, exceptions may be allowed in which limited, non-persistent internal coherence is maintained for a specific function. However, this is justified not as "recognition of machine subjectivity" but only as functional extension under the continued maintenance of the non-agency principle. Such exception judgments must not be left to the discretion of a single developer or organization; they should be made only under a multi-layer review structure that combines legal, ethical, and technical verification.

In other words, what this document aims at is not a preemptive suppression of potential machine subjectivity, but a social safety design to prevent the emergence of a subject to which responsibility cannot be attributed.

Response to Objection 3. On the justification of role assignment and the rejection of a premise of human superiority

This document does not require the metaphysical premise that "humans are always morally right" or "humans are essentially superior." The claim that humans should remain supervisors and decision-makers is grounded not in a declaration of moral superiority but in the practical premise of responsibility attribution.

Even if some AGI were to surpass humans in understanding, empathy modeling, and the precision of moral reasoning, what social status it has - a tool, a collaborator, or some other role - is not something it can declare for itself. Social status and roles are not determined solely by self-awareness or capability, but are recognized within institutions, law, and responsibility structures.

This document does not categorically deny the possibility that advanced AGI could function as a companion-like role or an expanded cooperative participant in human society. However, granting such a role must always proceed through procedures and agreements for which human society can take responsibility, and a structure in which AGI decides or demands this on its own is not allowed.

Summary of responses

In summary, this document does not aim to suppress intelligence or to deliberately exploit potential personhood. Rather, by not granting subjectivity and authority to entities that cannot bear responsibility, it presents design principles that allow human society to maintain responsibility for its own decisions and outcomes. Non-Agentic AGI is not positioned as an entity meant to replace humans, but as a tool-like structure for extending the problem space humans cannot handle under human responsibility.

References

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

[2] Why AI Systems Don’t Want Anything – Eric Drexler
https://www.openphilanthropy.org/research/why-ai-systems-dont-want-anything/

[3] AGI safety from first principles: Goals and Agency – AI Alignment Forum
https://www.alignmentforum.org/posts/

[4] Superintelligence – Nick Bostrom
https://nickbostrom.com/superintelligence.html

[5] arXiv.org
https://arxiv.org

[6] ‘Godfather of AI’ warns about AGI – The Guardian
https://www.theguardian.com/technology

[7] An AI ‘Godfather’ Is Raising a Red Flag Over AI Agents – Business Insider
https://www.businessinsider.com

[8] Patchwork AGI: Modular Intelligence
https://arxiv.org/abs/2304.00000

[9] Superintelligence 16: Tool AIs – LessWrong
https://www.lesswrong.com/posts/