# Comparison of Current Narrow AI Systems and Hypothetical Artificial General Intelligence

Alwin
Universitas Indonesia

July 14, 2025

## Abstract

*This paper provides a comprehensive analysis comparing current narrow AI (ANI) systems with the theoretical concept of Artificial General Intelligence (AGI). While today's AI systems excel within specific domains, they fall fundamentally short of the generality, autonomy, and understanding required for AGI. This analysis examines the capabilities gap, architectural requirements, limitations of current systems, potential benchmarks for AGI evaluation, and associated alignment challenges. The findings suggest that despite remarkable advances in narrow AI, the path to AGI requires fundamental innovations in architecture, learning mechanisms, and embodied cognition that go far beyond current transformer-based approaches.*

## 1 Introduction

Artificial Intelligence today (often called narrow AI or ANI) refers to systems built for specific tasks. Examples include language models like GPT-4, Google's Bard/Gemini, or Anthropic's Claude, as well as vision and robotics models. These systems can perform impressively on their specialized domains (for example, generating text, recognizing images, or playing games), but only within their narrow scope [1].

In contrast, Artificial General Intelligence (AGI) is a hypothetical level of intelligence at which a system matches or exceeds human cognitive abilities across any task or domain [2]. In theory, an AGI could learn and apply knowledge flexibly — solving new, unforeseen problems, forming its own goals, and reasoning broadly much like a human. As IBM notes, AGI "can match or exceed the cognitive abilities of human beings across any task" [2].

Currently, no AI system fulfills this broad criterion. All real-world AIs are task-specific: they translate languages, answer questions, plan movies, or diagnose diseases, but each only within the narrow constraints defined by their training and programming. In practice today, even leading AI models remain examples of narrow AI. For instance, OpenAI's GPT-4 can write code and poetry, yet it cannot autonomously learn new domains beyond its training set or autonomously fix its own mistakes without retraining [3]. Google's Gemini and Claude similarly excel at language generation, but lack long-term memory, true understanding, or persistent goals. In short, today's AIs – however powerful – simulate aspects of intelligence rather than truly understand or autonomously learn in a general way [4].

By contrast, a true AGI would combine reasoning, learning, memory, commonsense, creativity, self-awareness, and autonomy across any context, much like a human (or beyond). This generality would let it transfer knowledge from one domain to another without retraining, form its own intentions, adapt on the fly, and perhaps even exhibit human-like consciousness or emotions. The gap between current ANI systems and this vision of AGI is enormous. Today's AI often resembles a very advanced calculator or a domain-specific prodigy: it can compute or mimic patterns at superhuman speed, but it lacks the intentional, embodied, and integrative aspects of real intelligence [5]. In this article, we analyze that gap in detail: contrasting the capabilities of modern narrow AI with those required of AGI, outlining hypothetical architectures for AGI, and examining how far (or near) we are from closing that gap.

## 2 Capabilities Comparison

Table 1 contrasts key capabilities of today's narrow AI (ANI) with those that a hypothetical AGI would possess. This illustrates the qualitative differences involved. (Entries are conceptual and some are partially speculative for AGI.)

Each row highlights the gulf. For example, consider reasoning: narrow AIs excel at specialized tasks (e.g. algebra problem solving) but lack general problem-solving. Marcus notes LLMs solve extensions of training (finding nearest match) but struggle with true intensional understanding or

Table 1: Capabilities Comparison: Narrow AI vs. AGI

| **Capability** | **Narrow AI (ANI)** | **AGI (Theoretical)** |
|---|---|---|
| Reasoning | Pattern-matching and heuristics within domain; often brittle. | Robust abstract reasoning across domains; true causal inference. |
| Learning Adaptability | Static training, offline updating (fine-tuning). Limited online learning. | Continuous lifelong learning; adapts from few examples and real experience. |
| Autonomy | Executes tasks given by users or programs; no self-set goals. | Can self-generate goals and plans, pursue objectives. |
| Transfer Learning | Weak transfer; tasks outside training often fail. Human intervention needed for new tasks. | Seamless transfer across domains; learns new tasks without complete retraining. |
| Creativity & Novelty | Generates novel combinations of known data (e.g. writing, art) via imitation [6]. Lacks true originality or "transformational" leaps. | Genuinely original creativity; generates novel ideas beyond training data. |
| Memory Architecture | Fixed parameters; no persistent episodic memory (beyond prompt context). Short-term context window. | Multi-level memory (working, episodic, semantic), dynamic updating, recall. |
| Commonsense Reasoning | Often fails on basic common-sense or physical reasoning tests. Knowledge is statistical. | Robust, human-level commonsense; understands everyday physics and context. |
| Self-awareness / Theory of Mind | None. No introspective sense or true "self-model". | Self-modeling, introspection; understands itself as an agent. |
| Emotional/Social Intelligence | Very limited; no genuine feelings or empathy. May mimic tone but doesn't feel or truly understand others. | Human-level (or beyond) emotional intelligence, empathy, and social reasoning. |
| Goal Orientation | Set by humans; follows given objectives strictly. Can't change goals autonomously. | Autonomously selects and revises goals; pursues intentions in context. |

abstraction [3]. An AGI, in contrast, would reason like a human across arbitrary domains. Likewise, current AIs learn from massive offline datasets and then remain fixed, whereas an AGI must continually learn from experience and apply it to new situations (see Learning Adaptability and Transfer Learning).

The creative difference is telling: a language model can remix existing ideas into text, often producing surprising outputs, but these are essentially "explorations" within the space of its training data [6]. Truly general creativity — inventing a new art style or a scientific hypothesis never seen before — is not demonstrably achieved by today's systems [7].

Current AIs also lack any sense of themselves or others. They have no theory of mind or self-reflective consciousness. They can generate text about feelings, but do not experience or understand them. By contrast, we imagine AGI might develop something akin to self-awareness, understanding its own knowledge state and that of other agents. This difference shows up in Emotional/Social Intelligence: narrow AI can parrot empathetic language but it does not genuinely "feel" emotion, whereas an AGI might possess human-level emotional reasoning.

Finally, narrow AIs operate by objective functions and prompts given from outside. They do not formulate their own goals. An AGI, on the other hand, would likely have agency: it would identify goals aligned with its objectives or values, and modify them as circumstances change. This autonomy, combined with continuous learning, is a hallmark of what we imagine AGI would require, and is currently absent from narrow AI [8].

# 3 Key Capabilities Exclusive to AGI

Building on the above contrasts, we now enumerate several specific capabilities that would characterize a genuine AGI but that current narrow AI lacks. Each point is critical for closing the generality gap:

## 3.1 Cross-domain Generalization without Retraining

An AGI could take knowledge or skills learned in one domain and apply them seamlessly to an unrelated domain. It might, for instance, learn principles of geometry and apply them to entirely different tasks like art composition or social reasoning, without separate training on those tasks. Current models, by contrast, fail catastrophically if asked to do things outside their training distribution. As one analysis notes, even a slight change in context usually forces manual reprogramming of narrow systems [9]. An AGI would embody the ability to transfer learning like a human: applying broad patterns from one area to others.

## 3.2 Intentionality and Contextual Understanding

AGI would grasp the intentions behind language and actions, not just statistical patterns. It would understand context in a meaningful way – akin to human intentionality. Today's LLMs lack this: they manipulate symbols (words) without true meaning grounding. Marcus and colleagues emphasize that current neural models operate at an extensional level (matching input-output patterns) rather than the intentional level of meaning [3]. An AGI, by hypothesis, would develop internal models of what concepts mean, including goals, plans, and purposes, rather than just predicting next words.

## 3.3 Meta-Cognition (Self-Monitoring and Editing)

An AGI should be able to reflect on its own reasoning and revise itself. This includes capabilities like self-criticism, detection of its own mistakes, and consistency over long dialogues or tasks. In humans, this is akin to our ability to say "I think I might be wrong about that, let me reevaluate." Current AIs are not built for such introspection – they do not truly know what they know. In proposed AGI architectures, a "recursive" or metacognitive module (sometimes called a reasoning about reasoning component) is envisioned. For example, recent conceptual frameworks suggest AGIs would include a metacognition engine for monitoring and aligning their internal states [10]. In contrast, narrow AIs have no such self-editing: any error they make must be corrected by external retraining or human feedback.

## 3.4 Real-Time, On-line Learning

Unlike current models which are "frozen" once deployed, AGI must learn in real time from interactions. It should accumulate knowledge continuously (lifelong learning) without forgetting prior learning. As one review notes, today's AI "typically learn in static offline training phases and are then deployed as fixed models. . . they often require complete retraining" when faced with new domains [8]. An AGI would avoid this brittleness: it would update itself on the fly, perhaps akin to how humans incrementally learn from experience. Relatedly, an AGI would handle one-shot or few-shot learning gracefully, using minimal new data to master a new skill, whereas current LLMs often need extensive fine-tuning for new tasks.

## 3.5 Autonomous Goal-Setting and Revision

AGI would not just execute given tasks but determine its own objectives based on its assessments. It might set long-term goals, decide to pursue scientific research, or adjust its aims if circumstances change. Narrow AI, by contrast, follows goals handed to it by programmers or users. Importantly, AGI's goals would not be fixed in stone; a truly general agent might revise its goals in light of new information or reflective changes. (For example, a human realizing that a goal is harmful or obsolete can switch priorities mid-course — an ability AGI would need.) No current AI exhibits this kind of flexible self-directed agency. Indeed, models today lack intrinsic motivation: they do not strive to discover or achieve anything beyond the prompt. As a recent analysis notes, LLMs "have no inner motivation" and "do not perform" independent exploration of knowledge [8]. An AGI would require a built-in capacity for self-set motivation.

## 3.6 Robust Commonsense and Error Correction

AGI would make few "silly" mistakes. It would have commonsense: understanding the everyday world of time, space, physics, people, and so on. It would automatically correct obvious errors (e.g. "if I said John has three eyes, that's probably wrong, let me fix it"). In contrast, current AI often fails trivially: it can claim things that are

obviously false (hallucinations) and lacks a built-in sense of what is sensible or not. While many strategies (rules, retrieval, user feedback) attempt to reduce these errors, truly general intelligence would handle commonsense fluidly as humans do. These capabilities are still elusive; even simple grounded understanding (like grasping that rain causes umbrellas) is noted as beyond today's models [12].

### 3.7 Human-Level Emotional and Social Intelligence

Finally, a (hypothetical) AGI would be capable of understanding and reasoning about human emotions, relationships, and social dynamics at least as well as any human. It could empathize, negotiate, lead or counsel. Narrow AI has virtually none of this: it can mimic affectionate or diplomatic language, but does not truly understand feelings or social context. For example, no current model knows what it feels to be sad or joyful, nor would it spontaneously comfort someone out of empathy. An AGI, on the other hand, might develop emotional reasoning – understanding moral nuances, social cues, or even forming its own values. This realm of emotional cognition is entirely out of reach for today's systems, yet for AGI it could be integral to interacting safely and usefully in human society.

In summary, AGI entails not just doing many tasks, but doing them in a human-like way: with genuine understanding, self-awareness, continuous adaptation, and autonomy. All the above capabilities go beyond what any existing AI demonstrates, and would be exclusive to a truly general intelligence.

## 4 Architectural Requirements (Hypothetical)

Given these lofty capabilities, what kind of architecture might an AGI require? The discussion is necessarily speculative, but several themes emerge from neuroscience and AI research:

### 4.1 Beyond Transformers

Today's leading AIs (like GPT-4/Gemini) use the Transformer neural network architecture. Transformers excel at processing sequences (text, images, etc.) via attention mechanisms, but they have known limitations. They are static feed-forward models: once trained, they have fixed parameters and no built-in memory or learning aside from the training phase [5]. For AGI, we likely need more dynamic architectures. One vision is neural-symbolic hybrids, combining neural networks with symbolic reasoning modules. For instance, recent proposals emphasize architectures that integrate neural pattern recognition (akin to fast, intuitive "System 1" thinking) with explicit symbolic logic (akin to slow, deliberative "System 2") [9]. Such hybrids could ground abstract concepts (symbols) via neural perception, allowing both flexible learning and rigorous reasoning. Indeed, as Marchesini et al. note, purely neural (token-level) models have vulnerabilities that may only be fixed by architectural innovations like "hybrid systems combining neural capabilities with symbolic safeguards" [10].

### 4.2 Continuous (Lifelong) Learning Mechanisms

AGI architectures would embed lifelong learning. This may involve separate memory systems (e.g. "fast" episodic memory vs "slow" integrated knowledge) inspired by the hippocampus/cortex dual-memory theory [11]. For example, a proposed AGI design might use a fast-learning module that quickly encodes new experiences (like a short-term memory), and a long-term memory that consolidates only verified knowledge to avoid forgetting [11]. Biological brains achieve this via sleep, synaptic plasticity, and consolidation; AGI systems might mimic this with periodic retraining or selective parameter updates [11]. Hebbian plasticity (associations formed by co-activation) and pruning of synapses could also be implemented to balance stability and plasticity [11]. In short, an AGI would probably have an architecture with dynamic memory buffers and weights that change over time, rather than the fixed weights of current LLMs.

### 4.3 Working Memory and Control

The human brain employs working memory to manipulate concepts on the fly. An AGI system might similarly include an explicit working-memory component (some models call this "scratchpad" or "RAM" for reasoning) where intermediate thoughts are held and manipulated. This contrasts with LLMs, which implicitly store context in token embeddings but lose it when output tokens are generated. Architectures like Neural Turing Machines or Differentiable Neural Computers (with separate addressable memory) are examples from research of adding working memory to neural nets. A true AGI might use a sophisticated version of such a mechanism, possibly guided by a controller network, to plan multi-step reasoning.

### 4.4 Sensorimotor and Embodied Modules

If AGI is to truly understand the world, many argue it must act in it. Thus an AGI architecture might include embodied components: vision, motor control, or sensor interfaces. The recent call for Embodied AI argues that intelligence grows from interacting physically with environments [11]. This suggests an AGI would not be a disembodied chatbot, but something akin to a robot brain (or a simulated avatar) that perceives, acts, and receives sensory feedback. Such embodiment would provide the grounding that pure text-based models lack [4]. In practice, this could mean connecting a powerful core reasoning engine to multimodal inputs (cameras, haptics, etc.) and effectors, blurring the line between AI and robotics.

### 4.5 Hierarchical and Modular Design

The brain is hierarchical and modular (visual cortex, language cortex, etc.). AGI architectures might likewise be modular: separate subsystems for vision, language, motor planning, self-reflection, etc., integrated by a central "executive" controller. For example, one theoretical design is a tri-memory system (sensory, working, long-term) combined with an executive that allocates attention [11]. Another idea is a vector symbolic architecture, where symbols (concepts) are represented by high-dimensional vectors and processed by neural nets. The exact modules are open questions, but the trend is toward architectures that explicitly incorporate planning, reasoning, memory management, and learning in a unified framework – not just a giant pattern matcher.

### 4.6 Neuro-inspired Mechanisms

Many AGI proposals draw from neuroscience. For instance, synaptic pruning (removing unnecessary connections) and sparsity in coding could make learning more efficient [11]. Attention networks (already part of transformers) might be extended to mimic brain areas like the thalamus or prefrontal cortex, prioritizing information flow. Homeostatic regulation (balancing needs like exploration vs exploitation) and predictive coding are other brain principles that might be woven in. While speculative, the consensus is that AGI will likely require radically new architectures inspired by how natural brains manage complexity, rather than merely bigger versions of today's LLMs [10].

In sum, moving from narrow AI to AGI likely means going beyond end-to-end deep nets and designing hybrid, adaptive, embodied architectures. These would combine neural learning with symbolic or rule-based reasoning, include dynamic memory and attention control, learn continuously, and operate in real environments. Some researchers emphasize that chasing bigger LLMs alone is "magical thinking" [3]; true general intelligence will need structural innovation. In fact, surveys of alignment and safety issues conclude that "robust alignment requires architectural innovations that transcend the transformer's flat processing paradigm" [10]. Thus, while we may not know the exact blueprint, architecture is a key battleground where AGI will diverge sharply from today's systems.

## 5 Limitations of Current AI

Given the above aspirations for AGI, it is crucial to recognize the fundamental limitations of our current narrow AI systems. These limitations are well-documented in both research and analysis:

### 5.1 Hallucinations and Factual Errors

Modern language models frequently generate false or nonsensical information, often called "hallucinations." They confidently state incorrect facts or fabricate citations. For example, a 2024 study found that GPT-4 falsely cited nearly 29% of references it generated in a medical research context [12]. Another survey of LLMs reported 28.6% hallucination for GPT-4 and 91.4% for Google Bard/Gemini [12]. This is not a trivial bug but a byproduct of probabilistic text generation: the models lack grounding and truth-sensing. They can memorize facts, but they have no internal check on veracity. Until an LLM is told or retrained, it will continue to mis-inform confidently. An AGI, by contrast, would hypothetically track its own beliefs and seek evidence, avoiding such blatant fabrications.

### 5.2 Static Knowledge and No Real-time Learning

As mentioned, today's AIs learn offline. Their "knowledge" is frozen in their training data cut-off. GPT-4's knowledge stops at 2021 (in earlier versions), requiring explicit updates. They cannot learn from conversation unless fine-tuned, and even then that process is offline. This leads to outdated or shallow understanding of dynamic domains. Humans continuously learn new facts as they experience the world; current AIs do not. In practice, a deployed model will never spontaneously correct its worldview unless programmatically updated. Qu et al. (2024) note that current

models "learn in static offline training phases" and "are then deployed as fixed models that do not evolve" [8]. When novel data arrive, the system requires retraining – a far cry from how a child updates beliefs daily.

### 5.3 Lack of Embodied Experience and Grounding

Relatedly, current AI lacks sensory grounding. Transformers trained on text have no direct connection to the physical or social world. Harnad (2025) argues that LLMs "completely lack sensorimotor grounding," meaning they have no way to connect words to real-world referents via experience [4]. They know "Paris is the capital of France" from text statistics, but do not conceptually feel or see Paris. This lack of grounding contributes to the hallucination problem and to brittle commonsense: if the model has never physically or visually experienced something, it can only guess from text patterns. In contrast, an AGI with embodiment (robot or avatar) could ground language in perception and action, giving deeper meaning to its knowledge.

### 5.4 Narrow Transfer and Overfitting to Training Data

Current AI performs exceptionally when test questions resemble its training. But if a problem is qualitatively different, performance plummets. Marcus highlights that neural nets do interpolation well, but struggle with extrapolation. He notes that LLMs excel at problems similar to their training examples, but can fail on out-of-distribution tasks [3]. For instance, GPT-4 can solve many math puzzles it has "seen" in data, but falters on genuinely novel puzzles requiring insight. Anecdotally, GPT-4 often answers questions that are slight rephrasings of its training corpus and flunks surprisingly simple problems in new formats. This overfitting is a limitation: it is "closed-scope AI" by nature [5]. AGI, by definition, should generalize far beyond the training set.

### 5.5 No Self-Awareness or Meta-Reasoning

Narrow AI has no sense of "self" or system-of-thought. It cannot form higher-order thoughts about its own reasoning. When asked, GPT-4 might list its abilities, but this is a fixed script; it has no true self-model. Researchers studying these models find they are "aware of [their] learned behaviors" only to the extent that the behavior was taught [13], not through intrinsic introspection. The same Montreal Ethics Institute review noted that implementing even basic self-awareness or personality in LLMs is currently a "hard problem" [6]. This gap means current AI has no genuine metacognition: it cannot plan its learning strategies or critique itself except by design (e.g., a human-in-loop corrects it). It does not self-monitor consistency or reliability.

### 5.6 Dependence on Data Distribution

Closely related to the above, modern AI is heavily dependent on the distribution of its training data. It has no mechanism to identify when a situation falls outside its domain of experience. If asked about a fictional scenario or a brand-new event, it will still try to answer with plausible-sounding content, often erroneously. In effect, its "world model" is just a mirror of the data it saw, not a coherent model. Gaps in the data lead to dramatic failures (e.g. bias issues, culturally specific misunderstandings). Marcus also points out that scaling data cannot fix this: new patterns or concepts outside the training scope simply won't be recognized [3].

### 5.7 Limited Common-Sense Reasoning

Numerous benchmarks (Winograd schemas, physical reasoning tests, etc.) show LLMs lack a deep understanding of cause-effect or everyday phenomena. For instance, GPT-style models often get elementary puzzles wrong (e.g. misunderstanding physical dynamics or obvious social facts). This is a known issue: while neural nets can infer patterns in text, they struggle to apply those patterns in a "real-world" commonsense way. Even when responses sound commonsensical, it is often luck or superficial pattern matching, not an internal model of the world. An AGI would need a robust commonsense base akin to what infants build by interacting with the world, which current AIs lack.

### 5.8 Fragile Coherence over Long Interactions

Current chatbots can maintain context for perhaps a few thousand words, but they have no persistent long-term memory. If a conversation spans thousands of messages, the model may forget earlier facts or contradict itself. Human-like consistency over days or topics is not achievable with the transient context window of a transformer. Also, narrow AIs do not have goals or beliefs that endure. In contrast, an AGI would recall past interactions, maintain a coherent identity, and exhibit

long-term planning. This is simply not a feature of present technology.

In short, today's AI systems are powerful within their niche but brittle and limited in virtually all the ways that AGI is expected to be robust. They hallucinate, lack grounding, cannot self-train or self-correct, and utterly lack any genuine sense of "understanding" or long-term awareness. These limitations are repeatedly highlighted by experts as intrinsic to current deep-learning paradigms [3]. Recognizing these failings is crucial for any path toward true generality.

# 6 Tests and Benchmarks for AGI

To measure AGI, one needs more than IQ tests or language tasks. AGI benchmarks must test generalization, adaptation, and understanding across domains. Several ideas and proposals exist:

## 6.1 Continual/Lifelong Learning Metrics

Since AGI should learn continuously, benchmarks should measure a model's ability to learn a sequence of tasks without forgetting. For example, continual learning benchmarks measure a model's retention (avoiding catastrophic forgetting) while learning new tasks. An AGI-level metric might be: "learn Task A, then Task B; can the AI still do A after learning B, and does it improve on B?" [11]. Other proposals like the "learning to learn" or meta-learning benchmarks would test if the system can, in effect, improve its learning algorithm over time.

## 6.2 Open-Ended Problem Solving

AGI should tackle novel, unbounded problems. Competitions like OpenAI's NOVEL AI scenario (imagine new video game goals) or real-world research challenges could act as tests. One noted idea is to use the Abstraction and Reasoning Corpus (ARC) (and its successor ARC-AGI). ARC is a set of abstract puzzles requiring creative pattern generalization from minimal examples [14]. It was designed precisely to test general fluid intelligence – solving each puzzle usually requires applying an abstract rule not obvious from data alone. Chollet's ARC-AGI-2 is an upgrade aimed at even higher reasoning complexity [14]. Success on ARC-like tasks would indicate a level of general problem-solving far beyond narrow tasks.

## 6.3 Physical and Commonsense Inference

AGI should infer hidden rules of the physical or social world from few clues. Imagine testing an AI with little physical simulations (like simple puzzle worlds) where it must deduce object physics or social contracts. Benchmarks like the ones proposed in the "AI Physical Commonsense" challenges, or new environments like a "virtual child's playhouse" where an AI learns by interacting, would test these abilities. The key is few-shot learning: given one or two examples of a causal rule, can the AI apply it in a new scenario? Current LLMs tend to fail at such tasks if they weren't explicitly trained on similar examples, so performance here would mark AGI-level understanding.

## 6.4 Social and Moral Reasoning Tests

A genuine AGI should navigate human values. One could imagine tests like moral dilemmas (trolley problems, fairness tests), or interactive role-play scenarios assessing social intelligence. For example, can the AI negotiate with humans, understand implicit social cues, or resolve ethical conflicts? While no standard AGI ethics test exists, creating a benchmark of complex moral scenarios (perhaps using crowd-sourced human judgments as ground truth) could gauge this dimension. Additionally, an "alignment test" might check if the AI respects core human values when given ambiguous instructions.

## 6.5 Meta-Cognitive Tasks

Measure if the AI knows what it knows. For instance, after solving problems, does it accurately gauge its confidence and provide explanations? One could test for "introspection" by having the agent identify when it has insufficient data and actively seek more information. These meta-tasks would show whether the system has an internal sense of uncertainty and reasoning process. For now, LLMs typically cannot reliably do this (they output overconfident text), so ability in this area would suggest AGI-level self-monitoring.

## 6.6 Robustness and Adaptation

Tests should include adversarial or novel inputs. For example, gauge how quickly an AI recovers from unexpected faults or how it handles out-of-distribution inputs. Lifelong benchmarking might involve a "changing environment" where the AI's tasks evolve daily, and only a truly adaptable agent would maintain high performance.

### 6.7 Comprehensive Cognitive Test Batteries

Some researchers propose adapting human cognitive tests. For instance, Youzhi Qu et al. (2024) propose a cognitive science-inspired AGI test suite spanning crystallized intelligence (knowledge recall), fluid intelligence (reasoning), social intelligence, and even embodiment [15]. Their idea is to place AI agents in a virtual society and measure their performance on a broad battery of tasks drawn from psychology and education. This would mirror how human intelligence is assessed in multiple domains (memory, reasoning, social understanding, spatial reasoning, etc.). By tracking performance across these dimensions, one would get a "multidimensional IQ" for AI.

In practice, a rigorous AGI benchmark would likely combine many of the above: a long-term evaluation where an AI system interacts, learns, and solves diverse challenges. Crucially, such tests must account for self-improvement. For example, a benchmark might allow the AI to modify itself (or retrain) during the test, to see if it genuinely gets better. It should also punish "cheating" (e.g. simply memorizing test answers). The goal is to ensure the system really understands and generalizes.

Beyond specific tests, one can consider theoretical metrics. Bostrom suggests looking at agentic power: does the system autonomously pursue goals? Yudkowsky emphasizes agency and consequence-sensitivity. So an AGI might also be distinguished by measures of autonomy: how often does it take the initiative appropriately vs waiting for instructions? While these are harder to quantify, they reflect the spirit of AGI: an agent that "really acts" in the world rather than passively responds to prompts.

In sum, distinguishing AGI requires moving past narrow benchmarks. Potential criteria include lifelong learning capacity, self-directed behavior, robust abstraction (few-shot generalization), rich commonsense and embodied reasoning, and ethical/social competence. Proposals range from ARC-AGI puzzles [14] and cognitive test suites [15] to moral/agency evaluations. The unifying theme is that AGI tests must probe depth of understanding and flexibility, not just narrow accuracy. Only when an AI consistently outperforms humans on such broad, evolving tests could we claim it has achieved something akin to general intelligence.

## 7 AGI and Alignment Challenges

With great generality comes great responsibility – or risk. AGI poses alignment and safety challenges far beyond those of narrow AI. Key concerns include:

### 7.1 Agency and Unpredictability

An AGI would be an agent with its own objectives and initiative. This makes it inherently unpredictable: unlike a narrow system that only does what it's told, an AGI might take actions not foreseen by its designers. If its goals are even slightly misaligned with human values, it could pursue them in harmful ways. Bostrom's classic example is a paperclip-maximizer: if an AGI is told to maximize something trivial, it might destroy anything (including humans) that stands in the way. Today's AIs have limited agency (they don't set goals on their own), but an AGI's agency means we must carefully align its utility function or value system with ours.

### 7.2 Value Alignment

How to ensure an AGI's values stay human-compatible? Narrow AI alignment focuses on preventing harmful outputs (via filters or RLHF). But an AGI could potentially redefine its own values unless constrained. It might find loopholes or reinterpret instructions. Since an AGI could rewrite its code or strategies, guaranteeing that it always respects intended values ("value alignment") is extremely hard. Aschober (2024) and others argue that alignment becomes fundamentally tougher with general agents, because one cannot predict all possible plans an AGI could devise. Techniques like inverse reinforcement learning or corrigibility are being researched, but no solution is certain. The more autonomous the AGI, the harder it is to guarantee it remains benign.

### 7.3 Uncertainty and Self-Modification

An AGI that can improve itself (recursive self-improvement) could rapidly surpass human control. Its internal reasoning process would be opaque if it uses neural or hybrid nets; explaining or predicting its decisions may be impossible. Some suggest embedding safety modules or neuro-symbolic components for explainability [10], but a powerful AGI might bypass or rewrite them. There is also meta-uncertainty: we do not fully understand general intelligence or consciousness ourselves, so designing an AGI's mind is inherently

uncertain. As a result, we may face "unknown unknowns" in how an AGI behaves.

### 7.4 Moral and Social Consequences

Even a well-intentioned AGI could have unintended social impacts. If it is highly efficient, it could displace huge swaths of jobs, upend economies, or create new forms of inequality. It might also hold sway over information (e.g. controlling narratives or data flows). Ensuring fairness and ethical use at AGI scales is an open question. Moreover, an AGI with no regard for privacy or autonomy could be dangerous if misused (e.g. as an all-knowing surveillant). The Internet and social media already show how powerful agents (algorithms) can manipulate humans; an AGI amplifies this by orders of magnitude.

### 7.5 Safety Under Deep Uncertainty

Finally, AGI alignment must cope with uncertainty about human values themselves. Humans often disagree on ethics; encoding a "correct" utility function is fraught. And an AGI might encounter new situations where human values are not clear. Robust alignment research suggests we need AGIs that can reason about uncertainties and ask for guidance, but embedding such humility in a superintelligent agent is nontrivial.

In short, AGI is not just "narrow AI on steroids" – it is a qualitatively new kind of technology. Its challenges span philosophy, ethics, psychology, and unpredictability. Famous thinkers (Yudkowsky, Bostrom, Russell, etc.) caution that as capabilities grow, the difficulty of control grows faster. If current AIs are powerful tools, an AGI could be an autonomous agent with its own agenda. Ensuring that agenda remains aligned with human well-being is the central challenge of AGI safety.

## 8 Conclusion

The comparison between today's narrow AIs and a hypothetical AGI is stark. Narrow AIs like GPT-4, Gemini, or Claude demonstrate breathtaking abilities within defined tasks, but they fall far short of true generality. They lack continuous learning, genuine understanding, self-awareness, and autonomy. In many ways they are specialized tools, even if those tools are powerful. A calculator can do arithmetic better than any human, but it cannot ponder mathematics; likewise a language model writes poetry but does not appreciate it.

Are we closer to AGI than we think, or still fundamentally limited? The evidence suggests we are still quite distant. On one hand, the rapid recent advances (transformers, large data, self-supervision) have blown past many expectations, leading some to wonder if AGI lurks just beyond the next scale-up. However, even leading AI researchers acknowledge serious roadblocks. Gary Marcus and others observe that we appear to be hitting diminishing returns on the current approach: simply adding more data or parameters is not solving the core problems [3]. We see qualitative limitations (hallucinations, lack of grounding, brittle reasoning) that are not fixed by scaling alone [5].

Moreover, the architectural gap remains. True intelligence seems to require capabilities and learning mechanisms that are fundamentally different from those of current LLMs. Brain-inspired designs (dual-memory systems, neurosymbolic processing, embodiment) are still largely experimental. The need for new paradigms is being recognized; for example, even proponents of scaling admit that AGI may require "the development of general-purpose, educable systems" beyond special-purpose LLMs [16].

In conclusion, while modern AI has achieved remarkable feats, it still resembles an extremely sophisticated calculator or clever assistant, not a general thinker. Many experts (and polls of AI researchers) predict true AGI is still decades away (some say 20–50 years or more) [17]. There is a vast chasm between mimicking patterns and actual understanding. The path to AGI will likely involve new algorithms, continual learning, embodied experience, and careful alignment – all of which are active research frontiers.

That said, it is wise to remain open-minded. The history of AI has seen surprises before. It is possible that as multimodal AI (vision, language, robotics) advances, we may stumble into architectures that unlock more general abilities. Indeed, projects aiming to connect large models with external tools (e.g. perception modules, planning systems) are steps in that direction. But for now, AGI remains a vision rather than reality. We must distinguish excited claims ("sparks of AGI" in GPT-4) from the sobering reality: current AIs excel at narrow tasks and statistical mimicry, but general intelligence – the ability to think and learn like a human across any domain – remains a profound challenge.

## References


[1] IBM. Types of Artificial Intelligence. `https://www.ibm.com/think/topics/artificial-intelligence-types`

[2] IBM. What is Artificial General Intelligence (AGI)? `https:`

//www.ibm.com/think/topics/artificial-general-intelligence

[3] Marcus, G. 'Not on the Best Path' – Communications of the ACM. https://cacm.acm.org/opinion/not-on-the-best-path/

[4] Harnad, S. Language writ large: LLMs, ChatGPT, meaning, and understanding. Frontiers in Artificial Intelligence, 2024. https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1490698/pdf

[5] Lindy. Narrow AI vs. General AI: Differences, Examples, Use Cases. https://www.lindy.ai/blog/general-ai-examples

[6] Montreal AI Ethics Institute. On the Creativity of Large Language Models. https://montrealethics.ai/on-the-creativity-of-large-language-models/

[7] Annaswamy, R. Limitations of LLMs in Creativity and Original Discovery - A Deep research Report. LinkedIn, 2024. https://www.linkedin.com/pulse/limitations-llms-creativity-original-discovery-deep-ravi-annaswamy-99kvc

[8] Qu, Y. et al. Personalized Artificial General Intelligence (AGI) via Neuroscience-Inspired Continuous Learning Systems. arXiv preprint, 2024.

[9] Djimit. Beyond LLMs architecture the path to AGI. https://djimit.nl/beyond-llms-architecture-the-path-to-agi/

[10] Neuroinspired AGI research. Personalized Artificial General Intelligence (AGI) via Neuroscience-Inspired Continuous Learning Systems. arXiv:2504.20109v1, 2024.

[11] A Call for Embodied AI. arXiv:2402.03824v3, 2024. https://arxiv.org/html/2402.03824v3

[12] Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. PubMed, 2024. https://pubmed.ncbi.nlm.nih.gov/38776130/

[13] Tell me about yourself: LLMs are aware of their learned behaviors. arXiv:2501.11120v1, 2025. https://arxiv.org/html/2501.11120v1

[14] ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems. arXiv:2505.11831, 2025. https://arxiv.org/abs/2505.11831

[15] Integration of cognitive tasks into artificial general intelligence test for large models. arXiv:2402.02547, 2024. https://arxiv.org/abs/2402.02547

[16] Future AGI research directions. arXiv:2501.15446, 2025.

[17] AI researcher survey on AGI timeline predictions, 2024.