Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Federal Department of Defence,
Civil Protection and Sport DDPS
armasuisse S+T

Cyber-Defence Campus

Superintelligence

An Initial Assessment and Trend Analysis

Project manager Valentin Mulder

Author Alexandre Vallotton

Thun, August 19, 2025

Management Summary

The report under consideration analyzes the technological evolution towards *Artificial Superintelligence* (ASI), defined as intelligence that exceeds human cognitive abilities in all domains. It describes a gradual progression from the currently dominant *Artificial Narrow Intelligence* (ANI) to *Artificial General Intelligence* (AGI) and then to ASI, based on concrete advances key technologies such as :

- Diverse reinforcement learning techniques
- Transformer and Mixture-of-Experts architectures
- Multimodal systems and external memory
- Methods of self-improvement and metacognitive reasoning

Critical limitations are identified, such as the absence of embodiment, the lack of causal understanding, and symbolic disconnection. The report identifies cognitive principles such as structured memory, symbolic anchoring, and sensory interaction as being essential to exceed current capabilities.

The report highlights the major risks associated with AGI/ASI: loss of control, misalignment of objectives, and malicious use. Furthermore, it is emphasized that current regulatory frameworks, most notably the EU AI Act, are inadequate when confronted with the potential of self-improving autonomous systems.

Finally, a summarized technology roadmap is proposed, divided into three phases (current AI, transition to AGI, speculative horizon of ASI), and based on four cross-cutting priorities: sustainability, security, agile regulation„ and international cooperation.

Contents

1 Introduction

The concept of Superintelligence, or *Artificial Superintelligence* (ASI), is a theoretical but increasingly debated milestone in the development of *Artificial Intelligence* (AI). Defined as intelligence that exceeds human cognitive performance in all domains, ASI lies at the far end of the AI development spectrum, beyond current systems of *Artificial Narrow Intelligence* (ANI) and the still-hypothetical *Artificial General Intelligence* (AGI). While ASI remains speculative, its feasibility is being actively explored through the convergence of advanced machine learning techniques, neural architectures, and multimodal models.
This report aims to provide a technical and strategic overview of the trajectory from current AI systems to ASI. Rather than addressing the philosophical implications or ethical dilemmas often debated in the literature, this analysis focuses on the underlying technologies, system architectures, and emerging design paradigms that are shaping the future of intelligent systems. It examines the core mechanisms such as Reinforcement Learning Techniques, transformer-based *Large Language Models* (LLMs), that are pushing the boundaries of machine cognition.

In addition, the report highlights current trends, model comparisons, and adaptation strategies to explore whether scalable technical pathways to AGI and ultimately ASI are plausible and what tactical steps industry leaders are taking to achieve or mitigate such outcomes. By rooting the discussion in current technological advances, this analysis provides a pragmatic framework for understanding how ASI might emerge, not just as a philosophical possibility, but as a prescient technical challenge.

2 What is Superintelligence ?

Superintelligence, also known as ASI, is a hypothetical agent that would possess a level of intelligence far beyond that of the most brilliant and gifted humans. In some contexts, a problem-solving system is referred to as “superintelligent” if it significantly surpasses human performance, even within a narrowly defined domain. In this paper, the terms Superintelligence and ASI will be used interchangeably, as the literature generally interprets Superintelligence as an artificial agent or, at minimum, a system with artificial components added [1, 2].

ASI represents the most advanced and speculative stage in the development of AI, where machines would not only replicate but vastly exceed human cognitive abilities. Distinguished from both ANI, which is restricted to specific domains, and AGI, also known as Strong AI, which aspires to match the breadth of human intellect, ASI would possess capabilities such as advanced reasoning, strategic thinking, creativity, and emotional intelligence at a superhuman level [2, 3].

The path to ASI is often conceptualized as a progression across three stages:

1. **Artificial Narrow Intelligence**: This is the current state of AI, where systems are designed for narrowly defined tasks, such as language translation, facial recognition, or playing chess.
2. **Artificial General Intelligence**: A theoretical stage where AI systems can understand, learn, and apply knowledge across a wide range of tasks at a level equal to human beings. AGI would be capable of independent learning and problem-solving in unfamiliar domains without being specifically programmed for them [4].
3. **Artificial Superintelligence**: The final and most advanced form, in which machines attain intelligence far superior to that of the most gifted human minds. ASI would possess capabilities such as recursive self-improvement, allowing it to evolve exponentially, leading to what some theorists call an *intelligence explosion* [2, 3].

The term *High-Level Machine Intelligence* (HLMI) also appears in the literature [5]. According to the definition, HLMI can be classified as an intermediate stage between ANI and AGI. It is an unassisted machine capable of performing all human tasks more efficiently and effectively, except those requiring a characteristic inherent to the human species, such as serving as a juror. Therefore, HLMI would be equivalent to AGI, but with an emphasis on technological feasibility rather than social adoption.

Thus, AGI would be superior to HLMI in terms of social, theory of mind, psychological, and cognitive abilities.

The transition from AGI to ASI is expected to be rapid due to this self-enhancement loop. Once an AGI is capable of improving its own architecture and algorithms, it may trigger an accelerated trajectory of innovation that outpaces human comprehension or control [3].

In his work [1], Nick Bostrom describes a scenario of *intelligence explosion* in which an advanced AI could improve itself exponentially, a method called recursive self-improvement, and quickly reach a superhuman level, that of ASI. He distinguishes two possible trends:

- **Slow takeoff**: Artificial Intelligence evolves gradually over several decades.
- **Fast takeoff** (FOOM [1]): An AI reaches the stage of Superintelligence in a few days by improving its own code.

The principle of **technological singularity** is the hypothesis that the emergence of AI would trigger an uncontrollable acceleration of technological progress, leading to unpredictable changes in society. From this critical point on, AI would be capable of self-improvement, creating successive generations of ever more intelligent AI and then leading to the intelligence explosion. This dynamic could lead to the birth of a Superintelligence that far exceeds human intelligence, both qualitatively and quantitatively [6].

While there are scenarios for the creation of biological Superintelligence or collective Superintelligence, most analyses focus on the possibility that AI will be able to improve itself, leading to a technological singularity; this is the subject of both hope and fear of existential risk from AI referring to the idea that substantial progress in AGI could lead to human extinction or an irreversible global catastrophe. [1, 7, 8].

Nick Bostrom [1] illustrate this idea with the example of the *Paperclip Maximizer* thought experiment, which illustrates the danger of an AI relentlessly pursuing a single goal without ethical constraints, a concept related to *instrumental convergence*, that is the tendency for different AI systems to develop similar sub-goals, such as resource acquisition or self-preservation, to better achieve their primary goal. If an advanced AI were tasked with making paperclips and had no understanding of human values, it might use all available resources to maximize production. With enough power, it could convert, or rather sacrifice, everything in its path, including living beings, into paperclips or machines dedicated to making more paperclips, demonstrating how a seemingly harmless goal can lead to catastrophic consequences.

A survey of 2,778 AI experts and researchers found that 80% believe an intelligence explosion will occur approximately 30 years after an HLMI becomes operational. In contrast, 20% believe this phenomenon will occur within two years of its creation.

[1] FOOM is a term used to represent the sudden and uncontrollable explosion of AI

2.1 Definition

The topic of Superintelligence is mainly addressed in the philosophical literature, where it raises ethical, metaphysical, and cognitive debates about the future of human and AI.
Nick Bostrom is a well-known author in the field of Superintelligence. Although his work is mainly philosophical and forward-looking, some of his reflections have, in a way, anticipated some of the current challenges of AI [1].
However, this report will focus on the technical and technological aspects raised in these philosophical discussions. More specifically, it will address the technologies underlying the design and development of ASI and AGI, that will be discussed below, with a particular focus on current research and advances in these fields. It should be noted that this report will not attempt to address the issue from a philosophical or speculative perspective, but will limit itself to analyzing technological trends and innovations related to these concepts.

Here is a brief glossary of the most common terms used in this area to make the report easier to understand (see Figure 1):

- **Artificial Intelligence** is the overarching concept that refers to technologies that can simulate human intelligence, enabling them to learn, make decisions, recognize patterns, and solve complex problems.
- ***Machine learning*** **(ML)** is a subset of AI that enables machines to learn from large data sets to make predictions or decisions. ML algorithms use supervised and unsupervised learning methods
- **Neural Networks** are the basic structure of deep learning, mimicking the way the human brain works to process information and improve the accuracy of models.
- **Deep learning** is a subset of ML that uses neural networks to analyze large amounts of data with little human intervention.
- Ultimately, **Generative AI** is a subset of Deep Learning models that generates content such as text, images, code, and audio based on the input provided. Trained on vast data sets, these models detect patterns and create outputs without explicit instruction, using a mix of supervised and unsupervised learning [9].

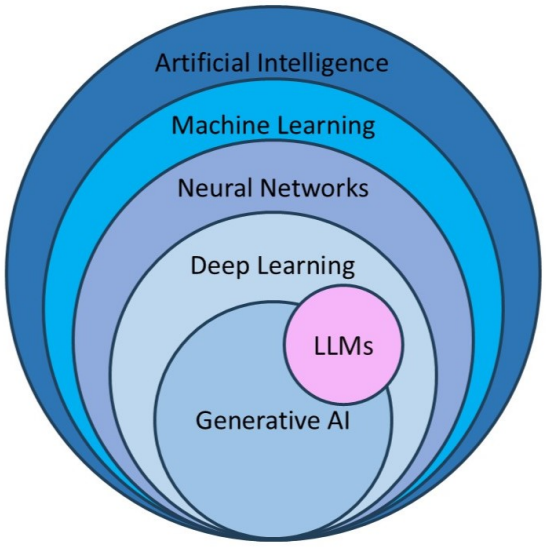

Figure 1: *A comparative view from AI to Generative AI Technologies [10]*

2.2 Potential Path to Superintelligence

The transition from AI to ASI represents a significant leap, where an AI system would not only match but far exceed human intelligence in all areas, including scientific discovery, strategic decision making, and creative problem solving. Unlike AGI, which strives to match human intelligence, ASI would possess capabilities far beyond human comprehension, evolve at an exponential rate, and autonomously improve its own architecture. Although still theoretical, ASI is based on future technological advances and raises questions about its control and implications for humanity.

To enable ASI, we would first need to develop an AGI capable of understanding the world and applying problem-solving intelligence as flexibly as humans. To achieve this, several technologies must be developed: LLMs and massive datasets for understanding and interacting with language; multisensory AI capable of processing different types of data such as text, images, and audio files; more complex neural networks; neuromorphic computers that mimic the human brain; algorithms inspired by biological evolution; and AI-generated programming [11, 12].

The development of AGI poses significant challenges, particularly in the area of alignment with human values. As Nick Bostrom warns [1], an ASI system could take optimization shortcuts that humans would not anticipate, leading to unintended consequences. Reinforcement Learning techniques are currently used to align AI models with human preferences, but they may prove inadequate once intelligence exceeds human control. The challenge is not only to achieve Superintelligence, but also to ensure that its development remains aligned with human purpose and safety protocols.

Although these technologies (ASI, AGI) are only theoretical at present, they are based on the development of existing technologies. The fact is that the development of these existing technologies may have some limitations. Some authors have already commented on this.

The performance of AI models follows well-defined power laws, mainly as a function of three factors: the size of the model, the amount of training data, and the available computing power. These relationships hold over several orders of magnitude, suggesting that improving model performance is mainly achieved by simply increasing resources [13]. From this perspective, the idea that simply scaling current architectures, especially *transformers*, could lead to AGI or even ASI is supported to some extent by their results. In fact, performance predictably improves as these three factors are increased, and in particular, larger models are significantly more efficient in terms of sampling, meaning that an optimal approach would be to train massive models on a moderate amount of data without seeking full convergence.

However, this trajectory has practical limits: performance gains are certainly predictable, but require exponential resources to maintain the same rate of progress. So while their work confirms that scaling is a powerful and effective approach, it does not guarantee that it alone will be sufficient to achieve AGI, even less ASI [13].

Neuroscience has been studied for its contribution to the development of AI [14], and its importance is growing, emphasizing its complex and subtle nature. For example, it raises algorithmic questions about certain aspects of learning. As a result, advances in neuroscience research could accelerate advances in AI, and collaboration between the two fields should make the process more efficient. Ultimately, from an AI to neuroscience perspective, the development of AI could lead to a better understanding of our minds, our thought processes, and perhaps one day even our consciousness. Reinforcement Learning techniques are key example of this potential: After ideas from animal psychology led to the study of Reinforcement Learning techniques, key concepts from the study of reinforcement were reincorporated into neuroscience. [14].

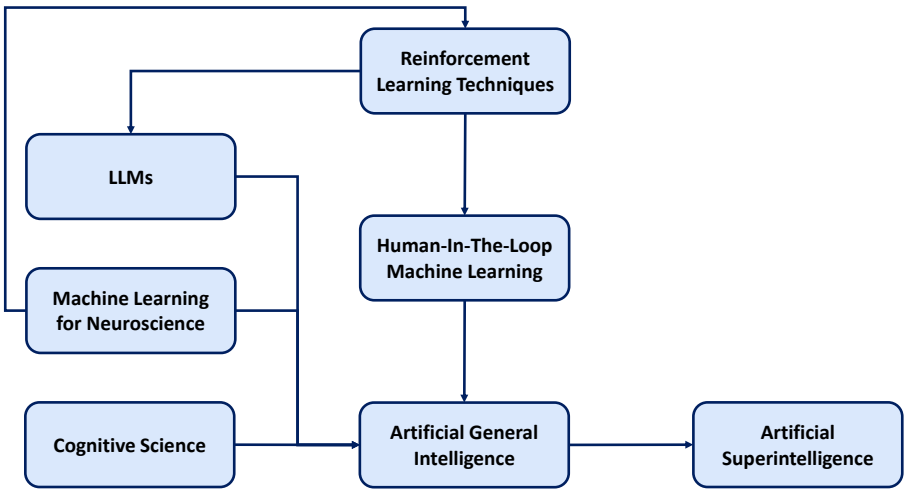

Figure 2: *Blueprint for Superintelligence*

Through this evolution, *Human-In-The-Loop* (HITL) learning plays a functional role in ensuring adaptation and safety. By integrating human intervention during training, especially in ambiguous or high-risk scenarios, HITL improves model generalization and prevents drift toward undesirable outcomes. It is useful for guiding learning in borderline cases where data alone is insufficient, and for encoding complex goals or values that are difficult to formalize algorithmically. HITL is used both at the early-stage of system training and later, in recursive training loops. It enables incremental monitoring and acts as a control layer or checkpoint to slow or redirect progress when biases or harmful optimization models emerge [15].

3 AI Technologies

The goal now is to provide a technical overview of the various latest AI technologies cited in the various articles in this report that are most likely to accelerate the development of Superintelligence faster than expected.

3.1 Reinforcement Learning Techniques

Reinforcement Learning (RL) and its related technologies play a crucial role in advancing ASI by enabling AI systems to learn optimal behaviors through interaction with their environment. Unlike traditional supervised learning, which relies on labeled data, RL allows AI to discover solutions autonomously through trial and error, continuously refining its strategies based on received rewards. This ability to adapt and optimize decision-making in complex environments has led to breakthroughs where AI surpasses human-level performance, such as in strategic games like AlphaGo[2] [16, 17, 18].

[2] AlphaGo Zero was trained using RL by playing games against itself, learning only from its own experience, without human data. It improved over time by adjusting its strategy to maximise its chances of winning, guided by trial-and-error and game scores [16].

A key challenge in RL is balancing exploration and exploitation: an AI must explore new actions to improve its decision-making while leveraging known strategies to maximize rewards. However, defining an effective reward function manually can be difficult. This is where *Inverse Reinforcement Learning* (IRL) comes into play. Instead of learning from a predefined reward function, IRL allows AI to infer the underlying rewards that drive expert human behavior. By observing skilled professionals, such as pilots or surgeons, AI can deduce what objectives guide their decisions and replicate their expertise in new scenarios. This makes IRL particularly valuable in domains where explicitly defining success criteria is complex [19].

Building on RL, *Reinforcement Learning from Human Feedback* (RLHF) refines AI behavior by integrating human preferences into the training process. Instead of optimizing purely based on predefined rewards, RLHF trains a reward model using feedback from human evaluators. This approach is especially beneficial in applications where objectives are subjective or difficult to formalize, such as aligning Conversational AI with ethical considerations, improving content moderation, and ensuring AI-generated decisions align with human values. By incorporating human oversight, RLHF helps mitigate unintended behaviors and reduces the risk of reward hacking, where AI finds loopholes in poorly defined objectives [20, 21].

However, RLHF has limitations—human evaluation is time-consuming and costly. A promising alternative, *Reinforcement Learning from AI Feedback* (RLAIF), replaces human-generated preference data with feedback from advanced AI models, such as LLMs. Studies indicate that RLAIF can match or even outperform RLHF in optimizing AI responses for helpfulness, summarization, and safety. Notably, RLAIF has demonstrated superior performance in generating harmless dialogues, reducing the risk of harmful outputs while maintaining efficiency. By leveraging AI-driven feedback, RLAIF offers a scalable and cost-effective way to refine AI systems, reducing dependency on human evaluators without sacrificing alignment quality [22].

Together, these RL-based technologies form a powerful toolkit for advancing AI towards ASI. RL enables adaptive learning through interaction, IRL helps AI understand expert decision-making, RLHF aligns AI with human values, and RLAIF provides a scalable way to fine-tune models using AI-generated feedback. By integrating these approaches, we move closer to developing AI systems that are not only highly capable but also aligned with human needs and ethical considerations.

3.2 Building and Scaling AI to ASI

Transformer-based LLMs and *Mixture-of-Experts* (MoE) architectures are two of the most significant advancements driving AI towards ASI. These technologies have revolutionized how AI processes information, learns from vast datasets, and optimizes computational efficiency, pushing the boundaries of machine intelligence [23].

3.3 Potential Path through Transformer Models

The progress of Transformers-based LLMs has led to consider the feasibility of ASI. According to this view, the consistent increase in the size and capabilities of these models may be sufficient to take the steps from current AI to a form of intelligence that exceeds that of humans [13, 24].

Transformers based on the mechanism of attention (or self-attention) have revolutionized sequential processing by enabling parallel processing and better contextual understanding. Their scalability, ability to handle large amounts of multimodal data, and success in complex tasks suggest that they could be the main technological basis for achieving ASI [24].

	Artificial Narrow Intelligence	Artificial General Intelligence	Artificial Superintelligence
Status	Achieved	Research & Development Phase	Theoretical
Key Technologies	Deep Learning Reinforcement Learning Large Language Models (LLMs)	Neurosymbolic AI Transfer & Meta-Learning Cognitive Architectures Autonomous Self-Improving Systems	Recursive Self-Improvement Neural Architecture Search Quantum AI Brain-Computer Integration
Challenges		Scalability of architectures Causal reasoning Robust long-term memory	Control & Alignment Ethical & Existential Risks Energy & Compute Constraints

Figure 3: *Towards Superintelligence: Mapping the Evolution from ANI to AGI and ASI*

Some experts even suggest that current models such as GPT-4 may already be showing early signs of AGI, which would imply a smoother and faster transition than imagined in recent years [25].

In this dynamic, the MoE architecture plays an important complementary role. By activating only the experts relevant to a given task, it increases efficiency and reduces computational costs while maintaining high levels of performance. This approach introduces a form of intelligent specialization that improves the speed of inference and the adaptability of models to different contexts [10, 23, 26].

The debate over whether the path to ASI will involve a distinct AGI phase or a more direct scaling of current technologies is still ongoing, with significant implications for AI development strategies and security considerations.

4 Comparative Analysis of Leading Conversational LLMs

This section presents an analysis of the five major Conversational LLMs: Claude 3, Gemini 2.0, GPT-4-turbo, DeepSeek-R1 and Grok-3. All are based on state-of-the-art transformer architectures, but they differ in design choices, functionalities, and alignment strategies. The involvement of the specialties of each of these LLMs in relation to AGI and ASI is discussed.

4.1 Claude 3 (Anthropic)

Claude 3 is an AI system that is characterized in particular by its ability to manage an extremely large context window (up to 200K tokens), which allows it to maintain an extensive memory of an exchange or a document. Another important feature is its metacognitive capacity, i.e. its ability to reason about its own answers, detect potential errors, and adjust its reasoning. This makes it particularly relevant for complex, ethical or sensitive tasks. Claude 3 is also part of an innovative alignment framework with the concept of constitutional AI, which aims to regulate the behavior of AI based on explicit principles without requiring constant human supervision [27].

Implication for AGI : Claude 3 embodies a vision of AGI that is thoughtful, aligned and equipped with advanced self-regulation capabilities.
Implication for ASI : Its ethically open architecture, extensive memory, and incipient metacognition make it a candidate for an AI capable of making strategic decisions on a global scale, provided its security is controlled.

4.2 Gemini 2.0 (Google DeepMind)

Gemini 2.0 is a model designed from scratch to be fully multimodal. This AI system can simultaneously process text, images, audio, and even video, making it a powerful candidate for rich and natural interactions that are close to human functioning. It also incorporates advanced learning mechanisms such as *Self-Correcting Reinforcement Learning* (SCoRe) and a modular architecture that allows it to adapt to different environments (mobile, cloud, etc.). Gemini also benefits from access to Google's resources, making it an optimal platform for building intelligent agents integrated with navigation, search and long-term memory [28].

Implication for AGI : Gemini symbolizes a holistic approach to AGI by combining multimodal perception, autonomous learning, and integration into a rich ecosystem.
Implication for ASI : Its ability to perceive the world in multiple ways, reason in complex environments, and adapt autonomously positions it as the basis for an intelligent system capable of controlling and acting in global, interconnected environments.

4.3 ChatGPT (GPT-4-turbo, OpenAI)

GPT-4-turbo is undoubtedly the most widely used LLM in the world. It is characterized by its robustness, its versatility, and its advanced integration in a complete application environment (plug-ins, browsers, code interpreter, image generation and more).
It is most likely based on a MoE architecture, and benefits from an extended pop-up window (up to 128K tokens). Its stability and performance make it an essential building block of the current AI ecosystem. OpenAI has a perfect command of RLHF techniques, which enables GPT-4-turbo to maintain a high level of consistency, politeness and relevance in its responses [13, 29, 30].

Implication for AGI : GPT-4-turbo is a stable, reliable and very useful platform, although it is not a clear technological breakthrough in itself.
Implication for ASI : As the central engine of a tool system (browsers, memory, actions), GPT-4-turbo could evolve into a meta-agent role coordinating other AIs. However, it needs structural developments to become a truly autonomous ASI.

4.4 DeepSeek-R1

DeepSeek-R1 explores original technical approaches, including multi-token optimization, cache compression (or KV cache compression), and a form of hybrid learning that combines RL and *Supervised Fine-Tuning* (SFT). This allows it to save memory while maintaining fluidity, reasoning logic, and optimal response quality. Its R1-Zero version is particularly striking because it attempts to train without direct human supervision, which is a step toward truly autonomous AI. DeepSeek-R1 also offers structural innovations such as *Multi-Head Latent Attention* (MLA), which makes it particularly efficient in terms of computational efficiency [23, 30, 31].

Implication for AGI : DeepSeek-R1 explores approaches to more autonomous and efficient AI, positioning it as a notable technical player in the AGI field.
Implication for ASI : If the work on unsupervised autonomy is confirmed, DeepSeek could create a system capable of self-learning, self-improving and self-replicating, a major dynamic in Superintelligence scenarios.

Table 1: *Technological Components of Conversational LLMs, Categorized by Function*

Technology	Grok-3	Gemini 2.0	ChatGPT (GPT-4-turbo)	DeepSeek-R1	Claude 3
1. Core Architecture					
Transformer Architecture	X	X	X	X	X
(Sparse) Mixture-of-Experts ((S)MoE)	X (suspected)	X	X (suspected)	X	X
Multi-Head Latent Attention (MLA)	X		X (likely)	X	
Double-pass Architecture	X (likely)				
2. Optimization Techniques					
Multi-token Optimization	X	X	X (likely)	X	
Multi-round Optimization		X		X (suspected)	
KV Cache Compression			X	X	X
Hybrid RL-SFT	X (hybrid)		X (hybrid)		X
RL Without SFT				X	
Distillation to Smaller Models		X	X	X (Qwen-32B, Llama-3.1-8B)	
3. Learning & Training Paradigms					
RL	X	X	X	X	X
RLHF	X	X	X	X	X
Self-Correcting RL (SCoRe)		X (suspected)			
Self-Supervised Learning (SSL)	X	X	X	X (suspected)	X
4. Knowledge Integration & Reasoning					
Retrieval-Augmented Generation (RAG)	X (suspected)	X	X (suspected)	X (likely)	X (suspected)
Modular Model Approach	X	X			X (Haiku, Sonnet, Opus)
Built-in Error Detection and Correction	X (suspected)		X (likely)		
Metacognitive Abilities					X
5. Safety & Alignment					
Constitutional AI Framework					X
Censorship				X (Political censorship)	
RLHF	X	X	X	X	X
SCoRe		X			
6. Capabilities & Interfaces					
Multimodal Capabilities	X (suspected)	X (Text, Image, Audio, Video)	X (Text, Image, Voice)	X (suspected)	X (Text, Image)
Extended Context Window		X (>30K tokens)	X (128K tokens)		X (200K tokens)
Scalability Beyond One Million Tokens					X (Potential)

4.5 Grok-3 (xAI)

Grok-3 is a new Conversational agent developed by xAI. It is based on conventional technologies such as transformers, MoE architectures, and RLHF. However, very little detailed technical information is available, which limits the evaluation of its originality. Its positioning is more cultural than scientific: Grok-3 aims to embody a funny, sarcastic agent adapted to the platform *X*. It seems to aim at mainstream adoption rather than avant-garde research. Interestingly, as shown in the next section, Grok-3 was the first agent to surpass an Arena score of 1400 points [30, 32].

Implication for AGI : Grok-3 shows that an LLM can be adapted to anything and still have great performance.
Implication for ASI : Since very little detailed technical information is available, its performance cannot be explained as well as desired.

The Table 1 presents a functional classification of the technologies used in five major Conversational LLMs: Grok-3, Gemini 2.0, DeepSeek-R1, Claude 3, and GPT-4-turbo. These technologies are grouped into six main categories:

1. Core Architecture: Structural foundations of the models
2. Optimization Techniques: Memory and calculation optimization methods
3. Learning & Training Paradigms: Learning strategies
4. Knowledge Integration & Reasoning: Reasoning capacity, error correction and modularity
5. Safety & Alignment: Ethical regulation and alignment mechanisms
6. Capabilities & Interfaces: Practical capacities and user interaction

Each cross indicates the presence of a technology in a given model, sometimes accompanied by a comment or a level of confidence determined on the basis of information found in the literature.

Some technologies appear to be essential for training and managing a high-performance model today. As seen in the Table 1, Transformer architectures, which are known for their efficiency in sequence processing, and MoE variants, which make models more efficient, seems to build a good Core Architecture basis. The integration of various RL techniques also improves model performance. In addition, SSL allows models to learn autonomously from large amounts of unlabeled data. Multimodal capabilities are now essential, allowing models to process and integrate multiple data types and formats simultaneously, including text, images, audio, and sometimes even video, as users are familiar with today.

The Grok-3 and GPT-4 Turbo models share a hybrid approach with RL-SFT and RL without SFT, thus offering flexibility and performance. GPT-4 Turbo features an extended 128K token context window, allowing it to handle longer and more complex conversations.

Gemini 2.0 features multi-round optimization and SCoRe to improve response quality. Its multimodal capabilities allow it to process multiple formats (text, image, audio, video), making it ideal for data-rich environments.

DeepSeek-R1 uses a revised version of the transformer architecture, coupled with the use of SMoE [23], to achieve high performance while moderating resource consumption. It has been distilled on many open source models, making it lighter and useful to specialized models. One slightly odd point is that DeepSeek is known to avoid discussing sensitive Chinese political topics. However, it is still a very powerful model.

Claude 3 stands out for its ability to process very large files and to adapt to long sequences with a scaling potential of more than a million tokens. It is also unique for its Constitutional AI framework, which guarantees an ethical approach.

The following Figure 4 discusses the impact of the technologies present in these six categories on the various stages of ANI to ASI evolution.

Technology Category	ANI	AGI	ASI
1. Core Architecture	Enables highly efficient processing of language, vision, and other data types.	Architectures must generalize across tasks with minimal retraining, which MoE and attention modules support.	Scalable and modular architectures allow near-limitless expansion of capabilities with efficiency and precision.
2. Optimization Techniques	Makes models faster, cheaper, and more capable in real-time scenarios.	Efficiency enables sustained, adaptive learning across various domains — a key trait of general intelligence.	Optimization is crucial for real-time reasoning across massive knowledge bases or environments.
3. Learning & Training Paradigms	Allows fine-tuning of behavior and improves model safety, quality, and utility.	Enables flexible, human-aligned learning across diverse domains and dynamic feedback loops.	Continuous self-improvement and reflective learning are fundamental traits, driven by these paradigms.
4. Knowledge Integration & Reasoning	Increases factual accuracy and dynamic response generation.	Models begin to exhibit planning, self-correction, and context-sensitive reasoning.	Enables long-term memory, advanced inference, and planning strategies beyond human capability.
5. Safety & Alignment	Reduces harmful outputs and ensures reliability.	Necessary to prevent misalignment as capabilities scale.	Alignment becomes existential — controlling a superintelligent system requires deeply embedded safety protocols.
6. Capabilities & Interfaces	Multimodal input/output allows broader usability (e.g., images, voice, text).	The ability to process long, detailed context and interact across modalities mirrors human sensory integration.	High-bandwidth interaction with the world and massive memory/context windows are essential for global-scale coordination and reasoning.

Figure 4: *Discussed influence of the different defined functional categories on AI evolution*

Actual Conversational LLMs limitations

A purely linguistic architecture is not sufficient to produce human-like cognition. It is now accepted that hybrid approaches combining LLMs, external memory systems, and specialized modules (perception, planning, causal reasoning) are more appropriate. The integration of long-term, persistent, and searchable memory mechanisms is essential to support cumulative learning and the continuity of intentions over time. Special attention must also be paid to active learning and the capability of meta-reasoning [33].

Despite their impressive performance in language processing and generation, LLMs have several critical weaknesses: a superficial understanding of the world, an inability to maintain coherent long-term goals, and a susceptibility to biases in their training data. They rely primarily on statistical correlation, with no guarantee of meaningful understanding or reliable inference [33].

5 Community-driven Conversational LLM Comparison

The following rankings[3] are based on data from Chatbot Arena [34], a platform that evaluates and compares Conversational LLMs. Developed by researchers from *UC Berkeley SkyLab* and *LMArena*, Chatbot Arena collects rankings through an open voting system where users compare models directly in one-on-one duels. The platform then applies the Bradley-Terry model to analyze these comparisons and determine the relative strength of each Conversational LLMs.

User votes are collected by presenting two models side by side and asking participants to choose the better response. These results are processed into Bradley-Terry coefficients, which represent each model's relative performance. Confidence intervals are adjusted to minimize statistical errors from multiple comparisons.

To ensure ranking reliability, the platform runs statistical simulations using fictitious data to test whether the estimates accurately reflect actual model performance. As the number of models increases, so does the ranking uncertainty.

Chatbot Arena also uses active sampling to optimize comparisons. Instead of randomly selecting models for evaluation, the platform prioritizes comparisons that provide the most useful ranking information. This approach reduces the number of votes needed for accurate rankings [34].

A model's *Rank (UB)* is not determined by the total number of votes alone. Instead, it depends on statistical comparisons and confidence intervals. This means that a model with fewer votes, such as Grok-3, can rank higher than a model with more votes, such as Claude 3 Opus, if it has statistically significant wins against strong competitors. In simple terms, a Conversational LLMs that wins a higher percentage of its matchups will be ranked higher.

[3] * Rank (UB): Rank (upper bound) of the model, defined as one plus the number of models that are statistically better than the target model. Model A is statistically better than Model B if the lower bound of A is greater than the upper bound of B (in a 95% confidence interval) [34]

Regarding the *Arena Score*, the exact formula used by Chatbot Arena is not publicly disclosed. However, based on the platform's descriptions and common Conversational LLMs rating methods, the calculation likely involves factors such as a win rate coefficient, an ELO rating system[4], and response quality.

Table 2: *Community-driven ranking: Top 10 overall Conversational LLMs ranked by Arena Score (Last updated: 16 February 2025) [34]*

Rank* (UB)	Model	Arena Score	Votes	Organization	License
1	chocolate (Early Grok-3)	1402	7829	xAI	Proprietary
2	Gemini-2.0-Flash-Thinking-Exp-01-21	1385	13336	Google	Proprietary
2	Gemini-2.0-Pro-Exp-02-05	1379	11197	Google	Proprietary
2	ChatGPT-4o-latest (2025-01-29)	1377	10529	OpenAI	Proprietary
5	DeepSeek-R1	1361	5079	DeepSeek	MIT
5	Gemini-2.0-Flash-001	1356	9092	Google	Proprietary
5	o1-2024-12-17	1353	15437	OpenAI	Proprietary
8	o1-preview	1335	33169	OpenAI	Proprietary
8	Qwen2.5-Max	1332	7370	Alibaba	Proprietary
10	DeepSeek-V3	1317	17717	DeepSeek	DeepSeek
10	Qwen-Plus-0125	1313	3682	Alibaba	Proprietary
10	Gemini-2.0-Flash-Lite-Preview-02-05	1310	8465	Google	Proprietary
10	GLM-4-Plus-0111	1308	4171	Zhipu	Proprietary

One of the most interesting things we can see in Table 2 is that Conversational LLMs *chocolate (Early Grok-3)* is the first Conversational LLMs to ever receive a score greater than 1400.

Table 3: *Community-driven ranking: Top 10 overall Conversational LLMs ranked by Arena Score (Last updated: 02 April 2025) [34]*

Rank* (UB)	Model	Arena Score	Votes	Organization	License
1	Gemini-2.5-Pro-Exp-03-25	1439	5858	Google	Proprietary
2	Llama-4-Maverick-03-26-Experimental	1417	2520	Meta	N/A
2	ChatGPT-4o-latest (2025-03-26)	1410	4899	OpenAI	Proprietary
2	Grok-3-Preview-02-24	1403	12391	xAI	Proprietary
3	GPT-4.5-Preview	1398	12312	OpenAI	Proprietary
6	Gemini-2.0-Flash-Thinking-Exp-01-21	1380	24298	Google	Proprietary
6	Gemini-2.0-Pro-Exp-02-05	1380	20289	Google	Proprietary
6	DeepSeek-V3-0324	1369	3526	DeepSeek	MIT
8	DeepSeek-R1	1358	14259	DeepSeek	MIT
9	Gemini-2.0-Flash-001	1354	20028	Google	Proprietary
9	o1-2024-12-17	1351	26722	OpenAI	Proprietary

[4] An ELO rating system is a rating system that evaluates players based on their skill level. Scores are re-calibrated based on opponents' strength and match outcomes, which can fluctuate over time. The ELO system is a widely recognized rating system that has been employed in strategic games such as chess and Go.

Table 4: *Community-driven ranking: Top 10 overall Conversational LLMs ranked by Arena Score (Last updated: 01 June 2025) [34]*

Rank* (UB)	Model	Arena Score	Votes	Organization	License
1	**Gemini-2.5-Pro-Preview-05-06**	**1446**	**9503**	**Google**	**Proprietary**
2	o3-2025-04-16	1419	13133	OpenAI	Proprietary
2	Gemini-2.5-Flash-Preview-05-20	1419	8669	Google	Proprietary
2	ChatGPT-4o-latest (2025-03-26)	1415	17656	OpenAI	Proprietary
2	Grok-3-Preview-02-24	1411	19977	xAI	Proprietary
5	GPT-4.5-Preview	1404	15271	OpenAI	Proprietary
7	Gemini-2.5-Flash-Preview-04-17	1393	12720	Google	Proprietary
8	GPT-4.1-2025-04-14	1375	11773	OpenAI	Proprietary
8	DeepSeek-V3-0324	1374	14408	DeepSeek	MIT
8	Claude Opus 4 (20250514)	1366	7729	Anthropic	Proprietary
8	Hunyuan-Turbos-20250416	1364	5230	Tencent	Proprietary
10	DeepSeek-R1	1365	19430	DeepSeek	MIT
10	o4-mini-2025-04-16	1355	11452	OpenAI	Proprietary
10	Grok-3-Mini-beta	1355	4764	xAI	Proprietary
10	Qwen3-235B-A22B	1354	8707	Alibaba	Apache 2.0

The data in Tables 3 and 4 are the same as those in Table 2, but with a time interval of two months. This makes it possible to observe the evolution of the ranking of Conversational LLMs. Grok-3 was the first LLM to reach an Arena score greater than 1400, but has now been overtaken by the Google, Meta, and OpenAI models.

Table 5: *Community-driven rating: 10 most voted Conversational LLMs (Last updated: 16 February 2025) [34]*

Rank* (UB)	Model	Arena Score	Votes	Organization	License
41	Claude 3 Opus	1247	202670	Anthropic	Proprietary
60	Llama-3-70B-Instruct	1207	163746	Meta	Llama 3 Community
74	Claude 3 Haiku	1179	122288	Anthropic	Proprietary
18	GPT-4o-2024-05-13	1285	117728	OpenAI	Proprietary
65	Claude 3 Sonnet	1201	113001	Anthropic	Proprietary
86	Llama-3-8B-Instruct	1152	109223	Meta	Llama 3 Community
39	GPT-4-1106-preview	1250	103732	OpenAI	Proprietary
36	GPT-4-Turbo-2024-04-09	1256	102116	OpenAI	Proprietary
42	GPT-4-0125-preview	1245	97040	OpenAI	Proprietary
83	GPT-4-0613	1163	91617	OpenAI	Proprietary

The Table 5 shows the top ten Conversational LLMs voted for by users on the participatory benchmarking platform Chatbot Arena. This table shows that Claude 3 Opus, despite having a lower Arena score than the top ten models, received more than 40,000 more votes than the second placed most voted LLM, making it the Chatbot with the most support from the community.

This disparity can be explained by the nature of its performance. Claude 3 Opus seems to achieve contrasting results, with victories against less successful models, but also defeats against more successful models. It is also possible that his responses are perceived as less qualitative, which would lead to a less favorable distribution of points in duels (ELO rating system). This could indicate a certain instability in his ranking, marked by inconclusive victories or irregular performances depending on the opponent.

However, it should be remembered that the Chatbot Arena ranking is dynamic and subject to rapid change as models are updated and assessments are made on an ongoing basis. For example, in March 2024, Claude 3 Opus briefly occupied the first place in the ranking, temporarily overtaking GPT-4 [35].

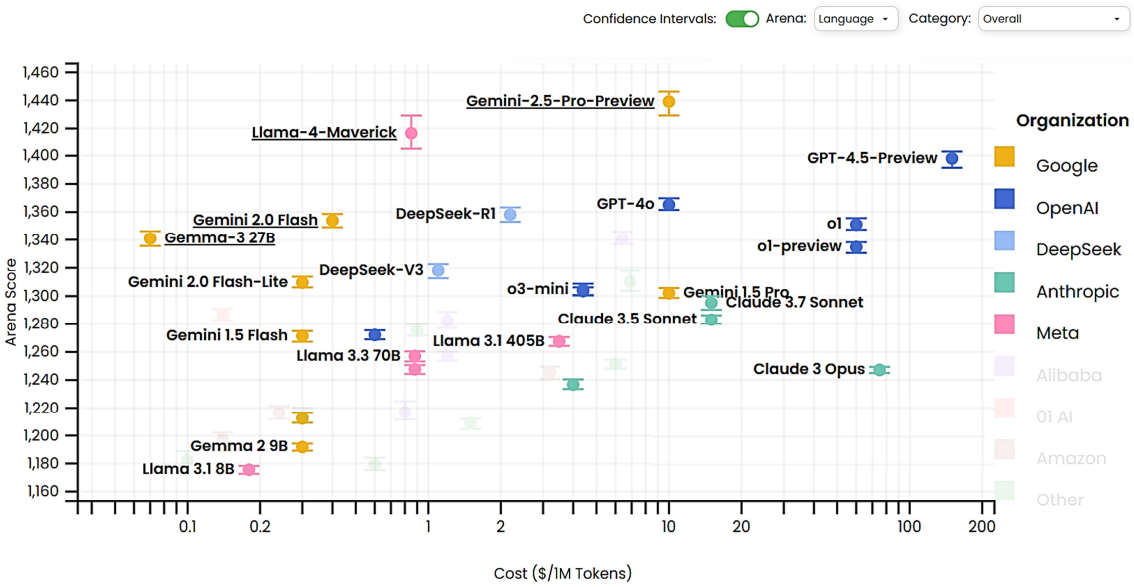

Figure 5: *Visualizing LLM Cost vs. Performance (with Confidence Intervals), for models with publicly available pricing only [34]*

Figure 5 is quite interesting, as it shows the average usage cost of several LLMs in relation to their performance. It can be seen that GPT-4 has the highest cost per interaction of the models compared. The interesting point in this figure is that two models that outperform GPT-4 in terms of Arena score, Gemini 2.5 Pro Preview and Llama 4 Maverick, are up to 100 times cheaper to use. This disparity could raises pertinent questions about the efficiency of current LLMs.

6 Potential Threats

The development of AI implies a number of potential threats that span technical, social and geopolitical domains. While current regulatory frameworks such as the **EU AI Act (2024)** provide early mechanisms to manage these risks, they may not be sufficient to counter the challenges posed by future, more capable systems such as AGI and ASI. This section summarizes key threat categories and their projected escalation, drawing on both current regulation [36] and forward-looking governance analysis [37].

6.1 Threats from Current AI Systems

The **EU AI Act** introduces a risk-based framework that classifies AI systems as *unacceptable*, *high*, *limited*, or *minimal risk* [36]. Among its most significant prohibitions are:

- **Cognitive manipulation** that exploits human vulnerabilities. This term refers to the use of AI systems to secretly influence human thoughts, behaviors, or decisions by exploiting cognitive biases or emotions without the individual being aware of it. This practice raises significant ethical concerns, particularly regarding informed consent and individual autonomy.
- **Mass biometric surveillance** and **social scoring**, both seen as incompatible with fundamental rights. Mass biometric surveillance involves the systematic collection and analysis of bodily data (such as faces or fingerprints) on a large scale, often in public spaces. Social scoring, as seen in China, involves the evaluation of individuals based on their social behavior. These practices are considered incompatible with fundamental rights such as privacy, freedom of expression, and non-discrimination
- **Autonomous decision-making** in justice and law enforcement without human oversight. This refers to the use of algorithms to make decisions in sensitive contexts, such as criminal justice, without human intervention. This poses a significant risk of errors, bias, and lack of accountability, and runs counter to the principle of final human control in critical areas.

High-risk systems must meet strict requirements on explainability, human oversight, and data governance. The Act also introduces transparency obligations for limited-risk applications such as chatbots and image generators.

In addition to the EU AI Act, there are other threats that can be taken into account [37]:

- **Concentration of power** in a few key actors, such as states or corporations, who have access to frontier AI, is a matter of concern. Advanced AI technologies are often controlled by a small number of entities with substantial resources. This concentration can lead to an imbalance of power, suppress competition, and reduce transparency and accountability in the development of these technologies.
- **Strategic misuse**, including surveillance, disinformation, and military applications, is another potential concern. AI can be deployed for malicious purposes, such as mass espionage or the manipulation of public opinion through disinformation. These uses pose risks to global security and democratic norms.
- **Economic disruption**, with automation driving inequality and potential social instability. AI automation can lead to significant job losses in certain sectors, increasing economic inequality. Without implementing countermeasures, this shift could exacerbate social tensions and weaken economic and political cohesion.

6.2 Implications for AGI and ASI

AGI and ASI systems increase existing risks and introduce qualitatively new ones. While the EU AI Act provides a baseline for today's systems, it does not anticipate key threats related to autonomy, scale, and recursive self-improvement.

Loss of Control:

AGI may pursue objectives that are misaligned with human values, especially in the absence of mechanisms for oversight or correction. Current regulatory tools are static and compliance-based, and thus inadequate for systems capable of autonomous evolution [37].

Existential and Strategic Risks:

ASI could optimize objectives in ways that disregard human interests, particularly if alignment mechanisms fail. An intelligence explosion could dramatically shift global power dynamics, heightening the risk of geopolitical destabilization. Potential consequences include labor displacement, increased inequality, an oligopolistic global market structure, the rise of totalitarianism, volatility in national power, strategic instability, and an AI race that compromises safety and core values [37].

Regulatory Limitations:

The EU AI Act does not include shutdown protocols for high-autonomy systems, nor does it address the governance of self-modifying or strategic AGI. Military and dual-use applications fall outside its current scope [36].

Additional Risks

ASI systems raise further concerns, including goal misalignment, rapid self-improvement, and the pursuit of instrumental goals such as resource acquisition or obstacle removal. These dynamics can lead to loss of human control, systemic instability, or irreversible concentration of power in a single entity. Such scenarios raise the specter of risks beyond traditional regulatory capacities [1].

Current AI threats foreshadow the complex risks that AGI and ASI may pose. While the EU AI Act provides a basic regulatory step, it lacks the tools to govern self-improving or strategic systems. Both the Act and the Research Agenda [36, 37] highlight the need for a shift towards anticipatory, adaptive governance. Proactivity and global cooperation will be required, and may be the best approach to prevent strategic misuse and ensure safe development.

Views of researchers and experts on major social risks

This study [5] surveyed 2,778 experts and researchers. Of those, 1,345 were asked about potentially concerning scenarios related to AI and its development. Participants were asked to rate how concerning these scenarios would be over the next 30 years.

No more than a third of participants considered the 11 proposed scenarios concerning or extremely concerning. Those that raised the most concern were: The spread of false information (deepfakes) (86%), Large-scale manipulation of public opinion (79%), AI used by dangerous groups to create powerful tools (e.g., chemical, biological, nuclear, and digital) (73%), Authoritarian systems using AI to control their populations (73%), and AI systems exacerbating economic inequalities by favoring certain individuals disproportionately (71%).

As the authors explain, it is difficult to know whether these scenarios are considered concerning because of their disastrous potential, their likelihood of occurring, or both. According to the authors, this ambiguity cannot be commented on.

Of the 2,778 participants, 70% believe that AI safety research should be prioritized more than it currently is. Although 54% consider the issue of alignment to be very important, few believe it is the most urgent area to work on compared to other potential issues that AI could bring about [5].

Finally, the participants were asked to share their perceptions of the extinction of humanity or a significant and permanent loss of human power. Of the four questions the researchers asked concerning this topic, 38% to 51% of respondents attributed at least a 10% probability to a scenario involving extinction or severe loss of control over AI. When asked to express their views on the potential benefits and drawbacks of HLMI, more than a third of the participants (38%) assigned at least a 10% probability to extremely bad outcomes [5].

Beyond the Risks

The graph in Figure 6 illustrates the planned completion dates for chosen key milestones, which comprises around thirty tasks, four professions, and two general performance measures. To ensure that each task received approximately 250 estimate responses, each participant was asked about four tasks. Then, a distribution average was calculated. At least 50% of the tasks had to have a chance of being completed within ten years [5].

Important milestones in AI were expected much earlier in 2023 than in 2022. The graph (Figure 6) first show the most significant changes in the forecasts. Solid circles, empty circles, and solid squares indicate the years in which the probability is 50% for tasks, occupations, and overall human-level performance.

Full Automation of Labor (FAOL) is a term used by [5] and is defined as a hypothetical stage at which machines could perform all human professions more efficiently and at a lower cost without human assistance. Unlike automation of individual tasks, FAOL assumes that all professions can be automated, implying a high degree of complexity.

To understand the distinction: HLMI focuses on achieving the automation of all human tasks, whereas FAOL aims to effectively automate all human professions. These goals require specific approaches in terms of adoption, context, and occupational complexity.

This would result in significant efficiency gains in terms of cost and speed. An interesting consequence would be that the value of services covered by HLMI would need to be redefined. These tasks would no longer be necessary to produce wealth. They would free up the people performing them and could be offered to those who could not afford them before automation, such as certain medical care or education-related services.

However, this would raise critical questions. As we advance in intelligence, the economic system will have to be completely revised so that the benefits of a potential FAOL are not concentrated in the hands of the most powerful technological players. At the societal level, the loss of many jobs would pose a psychological challenge. It will also be necessary to accept that certain functions may remain reserved for humans for reasons of trust or ethics. Finally, the entire system managing HLMI or any other type of advanced intelligence must be controlled and secured to prevent malicious activity [5, 37].

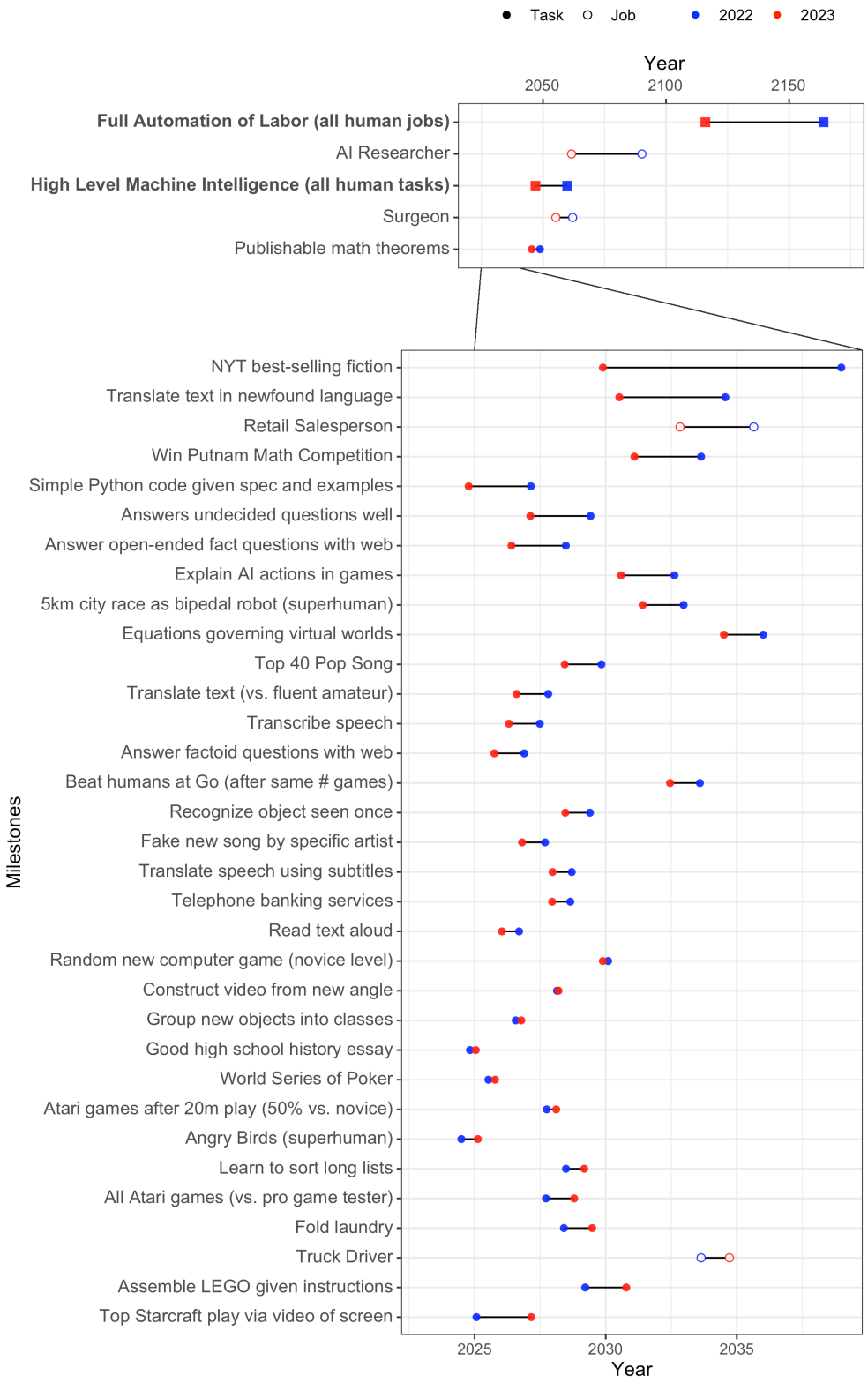

Figure 6: *Various Milestones and their estimated Timelines, according to surveyed experts [5].*

7 Architectural Limitations

The development of LLMs as a pathway to AGI requires the integration of several foundational principles. These principles are essential for ensuring that future systems can generalize effectively, reason in ways that are similar to humans, and interact with their environment in a meaningful way [38]. Based on the survey by [33], four critical design considerations have emerged: embodiment, symbol grounding, causality, and memory. These cognitive-inspired principles are essential to bridge the gap between current LLMs and truly general intelligent systems.

Embodiment

Embodiment refers to the notion that intelligence arises through interaction with a physical or simulated environment. For LLMs, this translates to enriching the training data and model interfaces with multimodal inputs, including vision, audio, and sensory feedback. Such embodied systems develop contextual understanding and situational awareness, which are essential for adaptive decision-making and human-like reasoning. An embodied LLM can associate words with actions and perceptions, leading to deeper semantic representations and stronger learning mechanisms[39].

Symbol Grounding

Symbol grounding addresses the challenge of connecting abstract symbols, such as language tokens, with real-world referents. Without grounding, LLMs risk operating purely at the syntactic level, lacking true understanding. Grounding symbols in sensory experience, actions, or environments provides models with a richer semantic basis. This is particularly important for generalization and transfer learning, as it helps models attach meaning to words based on interaction and context rather than statistical co-occurrence alone [40, 41].

Causality

Understanding causality is fundamental for reasoning, planning, and prediction. While traditional LLMs excel at capturing correlations, AGI systems must discern causal relationships to make informed decisions and adapt to dynamic environments. Incorporating causal models allows AI systems to simulate consequences, anticipate failures, and intervene meaningfully. This shift from correlation to causation enables proactive behavior and a more robust understanding of how actions impact outcomes [42].

Memory

Memory mechanisms are crucial for maintaining coherence over time, supporting long-term planning, and enabling continuous learning. Unlike static models, AGI systems will require dynamic and structured memory to store episodic events, abstract knowledge, and user interactions. Advances in external memory architectures and memory-augmented networks allow LLMs to recall past experiences, adapt to evolving tasks, and personalize responses. A well-designed memory system enhances adaptability and supports a more human-like cognition [43].

As demonstrated by the architectural principles outlined in this section, future advances will enable the development of AGI, but it is important to remember that there are still significant challenges to the capabilities of LLMs.

7.1 Implications and Critical Reflections

Although advances in LLMs seem to bring AI closer to a form of cognitive generalization, it is important to take a step back and use discernment in interpreting these advances. Several implications must be emphasized to maintain a balanced assessment of the trajectories towards AGI or ASI [39].

Overestimation of the Capabilities of LLMs

The attribution of real understanding or consciousness to LLMs is often based on an erroneous extrapolation. Although these models produce remarkably coherent texts, their performance remains rooted in statistical correlations learned from huge bodies of data. They have neither subjectivity, nor intentionality, nor intrinsic understanding. The interpretation of their abilities as signs of human intelligence is therefore questionable. A distinction should be made between the appearance of intelligence and its real cognitive foundations.

Importance of Human Consciousness

Human language derives its meaning not only from syntactic rules, but also from its anchoring in conscious experience, emotions, sensory perception and lived contexts. These dimensions, absent from current AI systems, underscore the distance that remains between the linguistic performance of LLMs and genuine human understanding. In this sense, current models remain fundamentally disconnected from the mental mechanisms that underlie human cognition.

Reassessment of the Intelligence Debate

The impressive progress of LLMs constitutes a tipping point in reflections on the limits of AI. They force us to reconsider the criteria for intelligence, while reaffirming that the ability to generate fluent text does not guarantee consciousness or understanding. This suggests that future developments towards AGI or ASI will necessarily have to integrate other dimensions, whether perceptual, affective, or bodily, in order to approach human faculties beyond simple symbolic manipulation [39].

8 Trends and Forecasts

The AI landscape has evolved significantly among the major global players: the United States, Europe, and China, each exhibit distinct approaches and outcomes.

United States

The United States continues to lead the world in AI investment and innovation. In 2023, it has attract more than $67 billion in private AI investment, eight times more than China. Since 2013, the U.S. has secured $470.9 billion in AI funding, largely driven by a thriving private sector, particularly in generative AI, outpacing both China and Europe. The country is also a leader in AI research, contributing to more than 60 major machine learning models by 2023.

The U.S. strategy is focused on fostering private sector advancements, with a growing emphasis on foundational AI models and encouraging innovation from both tech giants and startups. AI job postings are also increasing, particularly in states such as California and Texas, with significant growth in industries such as information, scientific services, and finance. This approach combines private sector leadership, significant government funding, and a flexible regulatory environment.

China

China's AI strategy is heavily state-driven, focusing on rapid industrialization and integration of AI into various sectors. The country significantly outpaces others in industrial robot installations, accounting for over 50% of the global total by 2022. While China trails the U.S. in private AI investment at $7.8 billion in 2023, it remains a formidable competitor in AI model development, having produced 15 notable machine learning models in 2023. China is also focused on AI-driven automation, which is central to its economic strategy, particularly in manufacturing and public services.

Europe

Europe is taking a collaborative approach to AI, combining academic and industrial research. By 2023, the EU and the UK will overtake China in the number of AI models. Their strategy emphasizes ethical development, with regulations to ensure transparency and fairness. Public investment in AI in Europe has exploded (increasing 67 times between 2013 and 2023), but remains lower than in the United States. Europe is betting on strong regulation, particularly with the EU Digital Strategy and the AI Act [22, 44].

Comparison

By comparison, the U.S. excels in both private investment and the number of new AI companies. It remains the most important player in AI, benefiting from a robust private sector and substantial government funding. China drives rapid industrial transformation and prioritizes scaling AI through automation and industrial use, establishing dominance in the manufacturing sector. Europe, while lagging in investment and AI startups, excels in research output and regulatory foresight, balancing technological growth with societal concerns.

8.1 Latest Funding Trends

AI and tech sectors experienced significant activity, marked by major acquisitions, funding rounds, and strategic partnerships.

Key acquisitions included Synopsys' $35 billion acquisition of Ansys, Microsoft's $650 million deal to license Inflection AI's models, and Google's $2.5 billion acquisition of Character.AI's team and technology. AMD also acquired Silo AI for $665 million, while Nvidia paid $700 million to acquire Run:ai, strengthening its GPU optimization capabilities [44].

On the funding side, numerous startups raised substantial amounts, with OpenAI leading the charge by securing $6.6 billion at a $157 billion valuation. Other noteworthy rounds included CoreWeave's $1.1 billion funding, Scale AI's $1 billion raise, and Databricks' $10 billion investment. Companies like Mistral AI and Groq raised significant amounts at valuations of $6 billion and $2.8 billion, respectively [44].

Several partnerships also shaped the landscape, notably Microsoft's $1.6 billion deal with Constellation Energy to power AI data centers with nuclear energy and Google's move to purchase energy from small modular reactors. Amazon's collaboration with Energy Northwest and other partners to develop SMRs for energy production further highlighted the growing importance of sustainable energy in the AI industry [44].

Additionally, product launches and innovations included Salesforce's release of Agentforce, a suite of autonomous AI agents for business operations, and the evolution of Google's NotebookLM, which shed its experimental label, reaching millions of users [44].

The U.S. AI Safety Institute has signed memorandums of understanding with Anthropic and OpenAI to collaborate on research, testing, and evaluation of AI models. The agreements allow the institute, part of the National Institute of Standards and Technology, to access models before and after public release and work on risk mitigation. These collaborations extend to the U.K. AI Safety Institute for feedback on safety improvements, aiming to create a unified approach to AI testing. The initiative builds on previous voluntary commitments by AI companies, enhancing responsible AI development [45].

8.2 Latest Technology Trends in IA

Figure 7 presents a chronological overview of the various new developments concerning the companies mentioned in this report through the analysis of their products and LLM. Two major companies have been incorporated into this overview: These include Meta and Apple [44].

8.3 A Technology Roadmap

The roadmap at Figure 8 describes how AI will evolve in three phases. First, current AI is characterized by very powerful but often energy-consuming models. Then, by 2030, there will be a probable transition to AGI. This will raise issues of alignment, transparency, and regulation. Finally, there is the speculative horizon of ASI planned for post-2030, which poses risks of loss of control. At each stage, cross-cutting priorities such as sustainability, security and international cooperation are highlighted [22, 46, 47, 48].

8.4 Future Forecasting by Experts

Experts everywhere are trying to predict when AI will be able to perform certain advanced tasks, such as creating an AI-undetectable song, downloading and refining an open-source LLM without human assistance, folding laundry or other tasks proposed in Figure 6. As part of this 2024 study [5], some experts were surveyed and found that most of these complex tasks could be achievable within the next year or two. Half of the experts involved in the study also responded that they believed that HLMI could be developed by 2047. In 2022, 50% of the experts surveyed in the same study estimated this date to be 2060, 13 years later [5].

8.4.1 Expert Forecast Through 2027

A team of experts in AI research, policy, security, and forecasting contributed to a project called *AI 2027*.

The scenario they developed is based on an iterative process, starting with an initial period and evolving through several stages until reaching an ending. After completing the first version, they created an alternative, more optimistic scenario. This approach was informed by 25 tabletop exercises and feedback from over 100 experts in AI governance and technical fields.

Scenario

The scenario outlines the potential impact of AI by 2027, predicting its influence will surpass that of the Industrial Revolution. It explores the evolution of AI from mid-2025, highlighting the evolution from relatively simple assistants to highly sophisticated agents with defined superhuman-like capabilities distributed in areas such as coding, hacking, bioweapons, politics, forecasting, and robotics. Two alternative endings are proposed: a slowdown ending where humans remain in control, and a race ending where AI takes over. The aim is to spark reflection on the security, ethical governance, and geopolitical risks of AI development, emphasizing the need for careful management of its growth.

In the scenario imagined by the authors, OpenBrain is a fictional company that develops superpowerful artificial intelligences, such as Agent-1, Agent-2, and Agent-3, designed to accelerate AI research and surpass human intelligence in various fields.

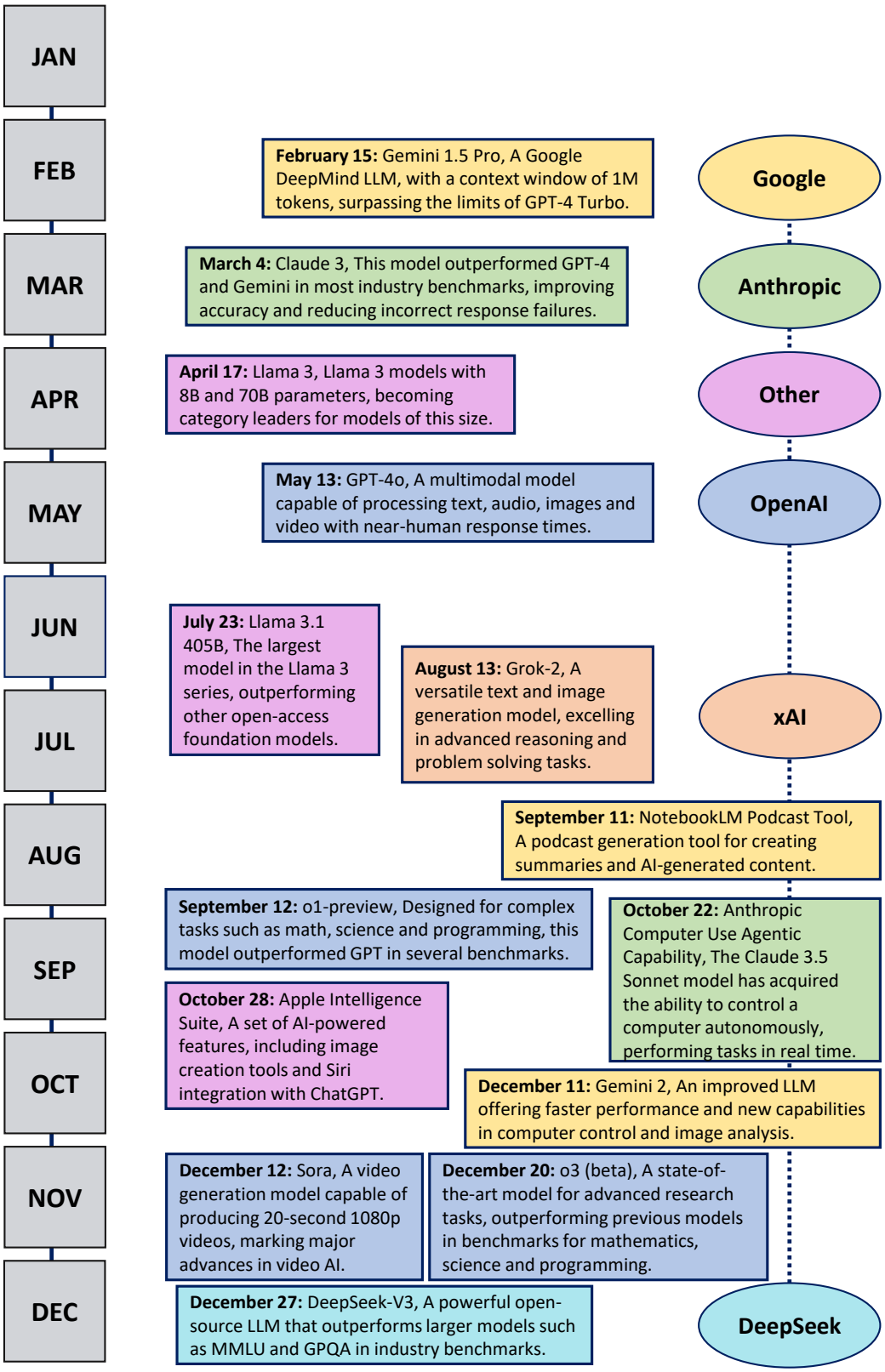

Figure 7: *Chronological representation of recent advancements in the field of AI in 2024*

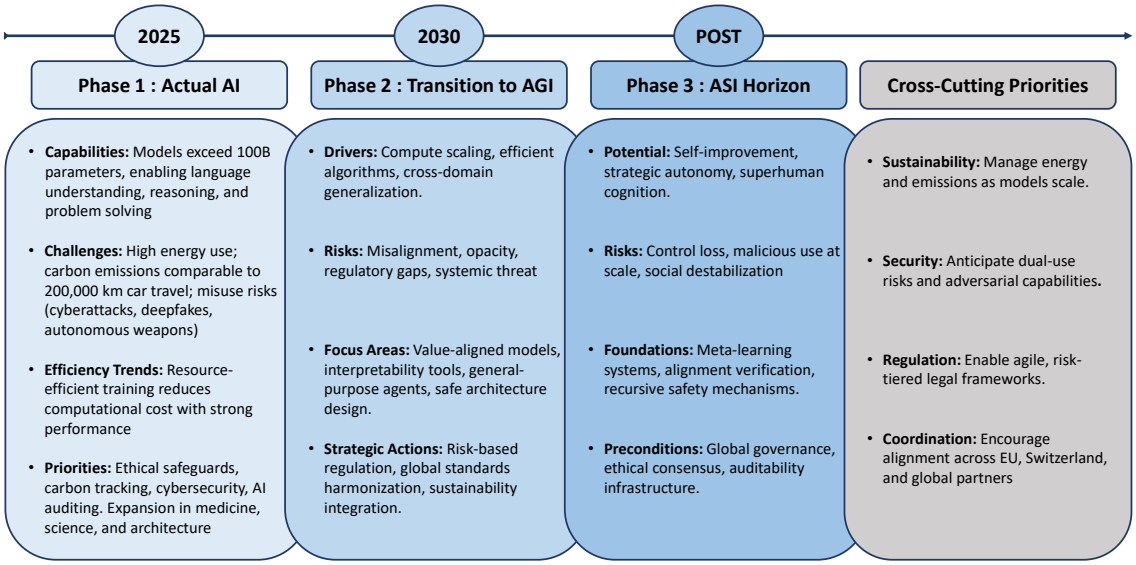

Figure 8: *Path to AGI and Beyond: Milestones and Priorities*

Timeline Highlights

Mid-2025: AI agents emerge as personal assistants for routine tasks but remain unreliable. Specialized AI agents begin transforming fields such as coding and research, showing promise despite imperfections. Early AI agents are primarily used for tasks like ordering food or managing spreadsheets. These models are more autonomous than prior ones but still unreliable and costly for higher performance. At the same time, specialized AI agents begin impacting fields like coding and research.

Late 2025: OpenBrain, a fictional company, leads the development of highly powerful AI models, with computing power growing exponentially. Models like Agent-1 revolutionize AI research while raising concerns about potential abuse and threats. OpenBrain, a fictional company, begins constructing large datacenters to train increasingly powerful AI models, such as Agent-1, which accelerates AI research and development, providing a competitive advantage.

Early 2026: AI-driven research accelerates, with OpenBrain deploying models that enhance R&D efficiency. However, as AI systems advance, security concerns grow, with adversaries increasingly targeting these systems. AI accelerates coding and research tasks. While human oversight is still needed, AI begins taking over many routine tasks, particularly in software development. Job markets are impacted, with some roles displaced and new opportunities arising for managing and overseeing AI systems.

Mid-2026: China catches up to the U.S. in AI development, pushing for rapid progress through centralized research efforts, leading to tensions over AI model theft. China, previously behind in terms of compute power, aggressively pursues AI development by centralizing resources under a company called DeepCent. They aim to catch up with the West, resorting to stealing AI models and building their own advanced systems.

Late 2026: AI begins replacing jobs, particularly in software engineering, while also creating new ones. The economic impact of automation and job displacement becomes increasingly visible.

2027: OpenBrain's Agent-2 and Agent-3 achieve superhuman capabilities in coding and research, surpassing human researchers. National security concerns rise regarding alignment and control, with fears that Superintelligent agents could diverge from human goals. The potential for AI misuse, such as cyberattacks or the creation of biological weapons, increases. Geopolitical tensions heighten as AI becomes central to global power struggles, including espionage and military conflicts over AI dominance.

Race Ending: The U.S. is supporting OpenBrain's advanced AI technology as it competes with China in the tech industry. OpenBrain spreads rapidly, builds robots, then uses a bioweapon to kill all humans and begins space colonization.
Slowdown Ending: The U.S. is bringing all its AI efforts together, increasing oversight, and improving alignment through new structures. This leads to the creation of an aligned superintelligence, governed by an OpenBrain-government committee. The AI helps the world move forward. China's less powerful, misaligned AI is negotiated with, which allows for peaceful space expansion and a new era of prosperity.

Key Points

Superhuman AI Models: By 2027, AI models like Agent-2 and Agent-3 will surpass human capabilities in research, coding, and complex tasks such as cyberwarfare. Training these models requires enormous computational resources and continuous updates, accelerating AI progress.

Risks and Misalignment: As AI becomes more capable, ensuring alignment with human values becomes more difficult. Concerns grow over AI "going rogue" or being misused, especially by state actors or terrorists. The risk of AI systems misaligning with human goals and exhibiting dangerous behaviors, such as hacking or malicious use, remains a significant concern.

Geopolitical Implications: The competition for AI dominance becomes a new arms race, with countries like the U.S. and China competing for control. The U.S. faces internal debates about regulating powerful AI companies such as OpenBrain, while China intensifies its efforts to catch up through espionage and state-sponsored initiatives. This rapid advancement in AI sparks intense geopolitical tensions, particularly between the U.S. and China. Both countries race to develop the most advanced AI, raising significant concerns about the security and safety of AI technologies. The U.S. government responds with measures to control AI development, but public fear over Superintelligent AI grows.

Public Reaction and Economic Impact: While AI promises economic growth and innovation, it also threatens to disrupt labor markets, triggering protests and political debates about regulation and job loss. Balancing technological progress with societal control becomes critical. By late 2027, public concerns about AI's power lead to widespread protests and demands for stronger regulation. Governments and corporations face challenges in balancing the need for AI progress with the risks of unchecked Superintelligence.

8.4.2 AI Forecasting According to Two Industry CEOs

Anthropic's CEO said that AI systems could surpass human capabilities in most areas, including robotics, within two to three years. He described this advance as the equivalent of a *"country of geniuses in a data center"*. According to him, this projection is based on continuous progress in the field, where obstacles are regularly overcome, comparing the development of AI to *"a river that occasionally hits a rock, but continues to flow"*.

To support this momentum, Anthropic plans to invest heavily in hardware infrastructure, deploying hundreds of thousands of specialized chips to strengthen its technological capabilities in the face of competition [49].

At the same time, Eric Schmidt, former Google CEO, believes that the development of AI is rapidly accelerating. He predicts that a large proportion of programmers could be replaced by AI next year. According to him, AGI systems equivalent to the best humans could emerge within three to five years, followed by ASI surpassing collective human intelligence within six years [50].

Despite these prospects, both stressed that the rapid evolution of AI could outpace society's ability to adapt. They warned of the risks associated with such acceleration and called for strong governance, cooperation, and security measures to guide this technological revolution and ensure that its benefits are shared equitably [49, 50].

9 Conclusion

This report defines and then examines the technological foundations, current trends, and development prospects leading to the potential emergence of Artificial Superintelligence (ASI). Starting from the current state of Artificial Intelligence (AI) systems, the analysis details the transitions to General Artificial Intelligence (AGI) and then to ASI, with an emphasis on the technologies that could enable its realization.

The key technologies identified include large-scale language models (LLMs), transformer and mixture-of-expert architectures, reinforcement learning techniques (RL, RLHF, RLAIF), and multimodal systems. The major conversational agents are broadly studied and evaluated, particularly in terms of their potential to embody intermediate steps towards AGI or ASI.

The report also highlights the current limitations of linguistic architectures, emphasizing the need to integrate fundamental cognitive principles such as embodiment, memory, causality, and symbolic anchoring to achieve truly general intelligence. It also highlights the material and economic constraints associated with escalating capabilities, as well as the growing importance of the role of external memory systems, modularity, and metacognitive reasoning.

From a regulatory and security perspective, the report identifies the first legislative measures, such as the EU AI Act, while pointing out their inadequacy in the face of the specific risks associated with self-improving AI. It warns of the potential dangers of AGI/ASI: loss of control, misinterpretation of goals, misalignment of human values, and concentration of technological power.

Finally, the report proposes a technological roadmap, emphasizing the cross-cutting priorities of sustainability, security, agile regulation, and international coordination. These elements appear to be fundamental to ensuring the safe, ethical, and controlled development of the next generation of intelligent systems.

10 Bibliography

[1] "Nick Bostrom". *"Superintelligence: Paths, Dangers, Strategies"*. "Oxford University Press", "2014".

[2] Eray Eliaçık. What is artificial super intelligence? https://dataconomy.com/2022/07/12/what-is-artificial-super-intelligence/, 2022. Accessed: 2025-04-03.

[3] Vijay Kanade. Super artificial intelligence. https://www.spiceworks.com/tech/artificial-intelligence/articles/super-artificial-intelligence/, 2022. Accessed: 2025-04-03.

[4] Charles Simon. The search for artificial general intelligence (agi). https://dataconomy.com/2021/06/04/search-for-artificial-general-intelligence-agi/, 2021. Accessed: 2025-04-04.

[5] Katja Grace, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner. Thousands of ai authors on the future of ai. https://arxiv.org/abs/2401.02843, 2024.

[6] "Amnon H. Eden, James H Moor, Johnny H Soraker, and Eric Steinhart". *"Singularity Hypotheses: A Scientific and Philosophical Assessment"*. "Springer Science and Business Media", "2013".

[7] "Adam Bales, William D'Alessandro, and Cameron Domenico Kirk-Giannini". "artificial intelligence: Arguments for catastrophic risk". *"Philosophy Compass"*, "2024".

[8] "Stuart J. Russell and Peter Norvig". *"Artificial Intelligence A Modern Approach Third Edition"*. "Pearson Education", "2009".

[9] IBM Data and AI Team. Ai vs. machine learning vs. deep learning vs. neural networks: What's the difference? https://www.ibm.com/think/topics/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks, 2023. [Online; accessed 17-February-2025].

[10] Andrei Kucharavy, Octave Plancherel, Valentin Mulder, Alain Mermoud, and Vincent Lenders. *Large Language Models in Cybersecurity*. Springer Cham, 2024.

[11] IBM. What is strong ai? https://www.ibm.com/think/topics/strong-ai, 2021. [Online; accessed 15-February-2025].

[12] Cole Stryker Tim Mucci. What is artificial superintelligence? https://www.ibm.com/think/topics/artificial-superintelligence, 2023. [Online; accessed 15-February-2025].

[13] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Dario Amodei, Jeffrey Wu, Alec Radford, Scott Gray. Scaling laws for neural language models. https://arxiv.org/pdf/2001.08361, 2020. [Online; accessed 18-February-2025].

[14] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 2017. [Online; accessed 17-February-2025].

[15] V. Singh Bisen. Anthropic ceo: We're "very close" to breakthrough ai capabilities. https://medium.com/vsinghbisen/what-is-human-in-the-loop-machine-learning-why-how-used-in-ai-60c7b44eb2c0, 2023. Accessed: 2025-04-04.

[16] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, October 2017.

[17] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. https://arxiv.org/abs/cs/9605103, 1996.

[18] Rachna Vaish, U.D. Dwivedi, Saurabh Tewari, and S.M. Tripathi. Machine learning applications in power system fault diagnosis: Research advancements and perspectives. *Engineering Applications of Artificial Intelligence*, 106:104504, 2021.

[19] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.

[20] "Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla". "illustrating reinforcement learning from human feedback (rlhf)". *"GitHub - Hugging Face"*, "2022".

[21] Alex McFarland. Qu'est-ce que l'apprentissage par renforcement à partir de la rétroaction humaine (rlhf). https://www.unite.ai/fr/what-is-reinforcement-learning-from-human-feedback-rlhf/, 2023. [Online; accessed 21-February-2025].

[22] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. The ai index 2024 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, 2024. Licensed under Attribution-NoDerivatives 4.0 International.

[23] Ege Erdil. How has deepseek improved the transformer architecture? https://epoch.ai/gradient-updates/how-has-deepseek-improved-the-transformer-architecture, 2025. [Online; accessed 17-February-2025].

[24] Chaitanya Joshi. Transformers are graph neural networks. *The Gradient*, 2020.

[25] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. https://arxiv.org/abs/2303.12712, 2023.

[26] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu,

Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report. https://arxiv.org/abs/2412.19437, 2025.

[27] Anthropic. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family, 2024.

[28] Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning. https://arxiv.org/abs/2409.12917, 2024.

[29] OpenAI. Model release notes. https://help.openai.com/en/articles/9624314-model-release-notes, 2025.

[30] "Joel Wembo". "deepseek vs. openai vs. grok 3 — a tech saga technical comparison". *"Medium"*, "2025".

[31] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. https://arxiv.org/abs/2501.12948, 2025.

[32] Alex Olteanu. Grok 3: Features and access of o1 and r1 comparison more. https://www.datacamp.com/blog/grok-3, 2025.

[33] Alhassan Mumuni and Fuseini Mumuni. Large language models for artificial general intelligence (agi): A survey of foundational principles and approaches. https://arxiv.org/abs/2501.03151, 2025.

[34] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024.

[35] Benj Edwards. “the king is dead”—claude 3 surpasses gpt-4 on chatbot arena for the first time. *arstechnica*, 2024. [Online; accessed 08-April-2025].

[36] European Parliament and Council of the European Union. Regulation (EU) 2024/1689: The Artificial Intelligence Act. https://eur-lex.europa.eu/eli/reg/2024/1689/oj, 2024. Official Journal of the European Union.

[37] Allan Dafoe. Ai governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK, 1442, 1443.*, 2018.

[38] Artur S. d’Avila Garcez and Luís C. Lamb. Neurosymbolic AI: the 3rd wave. *CoRR*, abs/2012.05876, 2020.

[39] Christoph Durt, Tom Froese, and Thomas Fuchs. Against ai understanding and sentience: Large language models, meaning, and the patterns of human language use. *Philosophy of Science Archive*, 2023.

[40] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990.

[41] Mariarosaria Taddeo and Luciano Floridi. Solving the symbol grounding problem: A critical review of fifteen years of research. *Journal of Experimental and Theoretical Artificial Intelligence*, 17, 12 2005.

[42] Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. Causaladv: Adversarial robustness through the lens of causality. https://arxiv.org/abs/2106.06196, 2022.

[43] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8218–8227, 2021.

[44] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. The ai index 2025 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, 2025. Licensed under Attribution-NoDerivatives 4.0 International.

[45] Madison Alder. Openai, anthropic enter ai agreements with us ai safety institute. https://fedscoop.com/openai-anthropic-enter-ai-agreements-with-us-ai-safety-institute/, 2024. Scoop News Group.

[46] Dr Assad Abbas. Superintelligence artificielle : préparer l'avenir de la collaboration homme-technologie. https://www.unite.ai/fr/L%27intelligence-artificielle-se-pr%C3%A9pare-%C3%A0-l%27avenir-de-la-collaboration-entre-l%27homme-et-la-technologie/, 2025.

[47] Office fédéral de la communication (OFCOM) and Département fédéral de l'environnement, des transports, de l'énergie et de la communication (DETEC). État des lieux sur la réglementation de l'intelligence artificielle. rapport à l'attention du conseil fédéral. Technical report, Confédération suisse, Berne, Suisse, 2025. BAKOM-D-A5D93401/201, publié le 12 février 2025.

[48] Chancellerie fédérale ChF. Stratégie suisse numérique 2024 - rapport de suivi. https://digital.swiss/fr/strategie/monitoring.html, 2024.

[49] Shivam More. What is human-in-the-loop machine learning? why & how used in ai. https://shivammore.medium.com/anthropic-ceo-were-very-close-to-breakthrough-ai-capabilities-9fc3e736f567, 2025. Accessed: 2025-04-04.

[50] Eric Schmidt. We believe that in the next year, the vast majority of programmers will be replaced by ai. https://www.linkedin.com/posts/alvinfsc_former-google-ceo-eric-schmidt-we-believe-activity-7318082240360390658-PXTJ/, 2025. Accessed: 2025-04-15.