AI in early 2026 – from fascination to maturity

Challenges on the Road to Better AI

Remigiusz Kinas *

remigiusz.kinas@gmail.com

Principal AI Researcher at Ingenix.ai, co-author of the Polish language model Bielik.

January 19, 2026

Abstract

This article synthesizes the state of AI development in early 2026 and argues that the main direction of progress is shifting from the scaling of quantitative models toward qualitative systemic maturity. I analyze thirteen research areas (including continual learning, knowledge compression optimization, test-time compute, world models, agency, and multimodality) and their implications for AI reliability and safety. The central thesis asks whether the key challenge of the coming years is further growth in cognitive capability or closing the gap between the increasing agency of models and the possibilities of their interpretation, diagnosis, and controllability. I demonstrate that the boundary of "better AI" is becoming not only data and architectures but also energy costs and computational infrastructure, which give development a geopolitical dimension. I conclude the essay with a discussion of the limits of anthropomorphization and how human subjectivity changes in a world optimized by algorithms with superhuman reasoning efficiency.

Keywords: AGI, continual learning, meta-cognition, world models, AI interpretability, scaling laws, agency, AI safety, Open Source.

*In this article, AI assisted with: translating complex texts and concepts into Polish (DeepL). The original was written in Polish and the Claude Code agentic system powered by Opus 4.5 model translated it into English (with manual corrections). AI was also used to check for typos and spelling errors. Gemini-3-Pro-Flash was used to edit some sentences for style. Overleaf with an LLM plugin served as the LaTeX editor – providing text style improvement suggestions. Research was conducted using AI-powered search engines. To understand source publications throughout the year, OpenAI ChatGPT (versions 4 to 5.2) and the Gemini 3 model family were used. Website was created by Claude Code vibe coding (some manual changes were performed too). All ideas for categorization, reflections, and topic development were conceived by a human and written by human hand.

Contents

1 Introduction

The prevailing paradigm based on exponential increases in computational power and training data volume is colliding with new barriers—both technological and physical. We are transitioning from a phase of fascination with generative capabilities (although I am certain that this year we will be shocked multiple times by the abilities of GenAI models, or by new breakthroughs in robotics) to a phase where reliability, energy efficiency, and systems' capacity for continuous adaptation become paramount. We are moving from an era of scaling into an era of research and optimization, increasing efficiency not through revolutionary actions but through the effect of continuous improvement.

However, it must be honestly noted that the thesis of the exhaustion of the "brute-force scaling" paradigm remains a subject of sharp debate in 2026. Voices such as Leopold Aschenbrenner's [1] argue that scaling laws have not slowed down at all, but have merely changed their "fuel"—from raw internet data to gigantic amounts of synthetic data and powerful computational resources devoted to inference itself (test-time compute). From this perspective, the path to AGI does not lead through algorithmic elegance, but through the construction of clusters worth hundreds of billions of dollars, which by sheer mass of silicon and energy push the boundaries of intelligence. We thus face a question: in the race for supremacy, will the "smartest" system win, or the one backed by the largest nuclear power plant?

It is no coincidence that I used the safe term "better AI" in the title, because in my opinion, 2026 forces us to return to the definition of Minimal AGI. Shane Legg from Google DeepMind [2] defines it as the moment when a system can perform any cognitive task that a human is capable of performing. We are no longer looking for a genius solving mathematical puzzles—we are seeking "mediocrity" that is universal. We no longer ask whether AGI is possible, but at what level of this spectrum we currently find ourselves. We recall the "failures" of major AI players who, while capable of solving complex mathematical problems, stumbled on simple tasks like "strawberry" (although this is not the best task for an LLM, it exposes its current AGI limitations).

The last few weeks and the transition into 2026 prompted me to reflect on the passing achievements of AI and the future—what the coming year might bring. The longer I wrote this article, the more questions, reflections, and doubts appeared in my mind. Are these challenges for the current year, or perhaps a roadmap for the next several or even dozen years? And yet there is so much talk about AGI these days. The biggest players in this field, companies producing well-known AI systems, are competing to predict the arrival date of superintelligence—in a year, five, ten, or never (there is no consensus on this matter). Do they know something that we ordinary mortals, AI users, do not see? Are the great American and Chinese labs hiding AGI from us—a superintelligence that will solve the world's greatest problems in the future? Or perhaps, anticipating the looming failure to achieve AGI in

the coming years, have they narrowed its requirements to solving a narrow range of tasks, such as being a smart chatbot at the level of an academic lecturer (a frequent slogan of AI creators—"a model at PhD level")? Then an even more important question occurred to me—what will human future look like in several or a dozen years?

I have no doubt, however, that the topic of AGI has become very fashionable recently. Often, it is not analyzed in depth. General statements are expressed, lacking broad analysis of the current state of AI. Opinions are presented without a specific definition of AGI requirements or the target state. New concepts like ASI (Super Intelligence) are defined, further obscuring the picture. In my opinion, this harms AI, harms the understanding of what it is and where we are heading. There is also rarely talk about the challenges, problems, and directions of today's artificial intelligence. Yet we see progress—most of us experience AI and benefit from this technology. This progress accelerates day by day. These are no longer small steps but giant leaps. I have a personal impression that at the beginning of 2026, each day brings greater changes than a week's progress in 2024. At the same time, according to the creators of LLM Arena, the best models remain at the top for an average of 35 days and drop out of the TOP5 within five months. Claude 3 Opus, introduced in March 2024, currently ranks 139th on the LLM Arena list.

The entry into 2026 brings hard evidence that the barrier of task complexity we pose to AI is systematically decreasing. Just look at the *FrontierMath Tier 4* [3] benchmark, considered a bastion of unsolvable mathematical problems. Of 48 extremely difficult tasks, as many as 14 succumbed to the power of OpenAI's flagship model, GPT-5.2 (Pro) (as of January 11, 2026). I am quite certain that more FrontierMath tasks will fall this year. I predict we will exceed 50% solutions.

Among the latest achievements, actually while writing this essay, we are being surprised by further news from the field of research mathematics. On erdosproblems.com, a case was recorded in which a model from the GPT-5.2 family led (in a loop with a human and formalization in Lean) to the resolution of several Erdős problems. Importantly, the success did not result from the "magical intuition" of a single prompt, but from a process. The barrier was no longer "computational power" itself, but controlling hallucinations and closing gaps in proofs through iterative criticism and support from formal proof processes. Even if some of these "solutions" are later reduced to finding results in the literature or clarifying ambiguous task content, the very fact that the *LLM* $\rightarrow$ *corrections* $\rightarrow$ *formalization* loop works in practice shifts the boundary of what we consider achievable by AI in 2026. We are on the threshold of a new industrial revolution. While the previous one replaced human muscles, the current one replaces brain functions.

Another evolution is observed in the *ARC-AGI-2* [4] benchmark. We started 2025 with scores at a few percent, only to close it with an impressive effectiveness of 54.2% (also OpenAI's model, GPT-5.2 (Pro)) and a proposal for a new test, ARC-AGI-3 [5]. Comparing models on some known benchmarks—GPQA Diamond, HMMT, AIME 2025, MMMLU—loses

discriminatory power (results above 90%). At most, these tests can serve as so-called sanity checks to verify training quality. They can confirm that the model has not degraded (regressed) and has not "forgotten" fundamental logical principles due to an engineering error.

And what about vibe? Vibe-coding and vibe-designing (music, graphics) are developing. I remember my first attempts at collaborating with coding agents. Usually, the time spent correcting the agent's errors was greater than writing from scratch. Perhaps it was my incompetence, lack of knowledge. However, the situation changed toward the end of the year. Subsequent solutions fulfill their purpose wonderfully. Maybe not perfectly, but noticeably better. They help not only professionals but also those who would never have attempted programming in their lives. Even orthodoxy coders who write "from scratch," like Linus Torvalds, mention real benefits from using AI to write Linux kernels. But that's not all. When OpenAI defines AI maturity levels, setting level five as the goal—*"Organizations: AI that can do the work of an organization"*—the first solutions for vibe business appear, such as Atoms.dev [6]. A development environment based on the "AI Team" concept. It allows transforming natural descriptions of ideas into ready digital products. Atoms builds a "virtual" development team in which autonomous units divide work into stages of planning, architecture design, full-stack code writing, and deployment and testing. All with minimal human involvement.

In this article, I have defined thirteen areas that, in my opinion, will determine further AI progress, but also what the future of human-technology relations will look like and thus what the world will look like. From the "data wall" and the need to go beyond human data, through challenges related to memory and meta-level reasoning, to building the foundations of future human-machine symbiosis. These are several directions for AI to cease being merely a static knowledge archive and become a dynamic, reasoning system capable of something more than clever reproduction. In this article, I reference the latest publications, defining the current state of issues and simultaneously the entry point into 2026. The points described will allow me to systematically track technology development on the road to better AI (AGI). However, there is no rose without thorns. Ideal AI and the desire to address most of these points will cause AI to become significantly stronger than humans. So what about safety, interpretability, the future world? That is probably a topic for an entirely different article. But let us keep this in mind.

2 Continual Learning and Adaptation

The most treacherous characteristic of contemporary models is not lack of "intelligence," but its uneven quality (jagged intelligence). How should we understand this? Extreme proficiency alongside surprisingly primitive errors. The true metric of 2026, as Demis Hassabis of DeepMind [7] mentions, will therefore be consistency and reliability, not additional benchmark points. How can we help AI with self-improvement (adaptation "on the fly" to new tasks) while experiencing many obstacles such as model drifting, catastrophic forgetting, contamination (how to evaluate new information), insufficient computational power, and method effectiveness?

One such option is continual learning [8] [9]. Continual learning (CL) is a paradigm departing from the static model of training on frozen datasets toward dynamic systems that evolve with incoming information. In the context of LLMs (Large Language Models), this challenge involves efficient adaptation to changing data distributions without the costly retraining of the model from scratch every time new facts, legal regulations, or user preferences appear.

According to the latest analyses, CL in the world of large language models is realized in two main directions:

- **Vertical continuity:** involves gradual transition from general model abilities to highly specialized competencies. This process includes three stages: Continual Pre-Training (CPT), Domain-Adaptive Pre-training (DAP), and Continual Fine-Tuning (CFT).
- **Horizontal continuity:** focuses on the model's ability to adapt over time and across different domains, allowing it to assimilate new trends and facts while retaining historical knowledge.

The key technological barrier remains so-called catastrophic forgetting. This phenomenon occurs when a model, while learning new information, overwrites parameters responsible for previously acquired skills, leading to a sharp performance decline in old tasks. CL solutions aim to create "targeted adaptation" mechanisms that are much more resource-efficient, allowing model updates at a fraction of the computational cost of full training.

The challenges before us, which are already being partially addressed in research, can be divided into four categories:

A. **Architecture** – architectural dynamics – is the ability to change the model's physical structure a condition for excellence and achieving AGI?

   Will we stay with MoE (Mixture of Experts) [10], allowing the model to dynamically select specialized "experts" (subnetworks) for each processed token? Or perhaps with *Mixture-of-Depths* [11], assuming that the model dynamically decides which tokens require full processing through transformer layers and which can skip them? Such a solution allows for "intelligent" real-time allocation of computational power. This

means transitioning from rigid processing of every data element in the same way to an architecture that learns to selectively engage its resources only where information complexity requires it.

Will we observe more radical discoveries in the near future, aimed at partial or complete "on-the-fly" architecture changes?

B. **Training** – how to dynamically change model capabilities after deployment using training techniques?

In 2025, the first concrete mechanisms for persistent model adaptation appeared. For example, Self-Adapting Language Models (SEAL) [12]. In this method, the LLM generates its own training data ("self-edits") and uses it to update its weights through a reinforcement learning (RL) loop. This allows the model to permanently learn new information without training from scratch. A step toward LLMs that actually update and adapt based on their own experiences (more on this in Chapter "3. Experience Beyond Human Data").

Another approach is BDH (Dragon Hatchling: The Missing Link between the Transformer and Models of the Brain) [13], which abandons the rigid division between training and inference phases in favor of Hebbian Learning mechanisms. Instead of relying solely on backpropagation, the system mimics biological plasticity. Neurons that jointly respond to a given stimulus strengthen their connections in real-time. This makes learning a natural byproduct of information processing, eliminating the "Groundhog Day" effect where the model forgets the interaction context immediately after it ends.

The end of 2025 brings further proposals in the form of the *Nested Learning* [14] paradigm. The presented *Hope* module (admittedly still in research phase) is a "self-modifying" system. Its innovation lies in breaking with the "illusion of rigid architecture," where the model (weights) is separated from a static learning algorithm (optimizers like Adam, SGD, etc.). In the Nested Learning approach, the optimization algorithm becomes part of the network itself. The *Hope* module operates in a nested loop: while base layers process data, a supervising module analyzes error dynamics and rewrites weight update rules for specific neurons in real-time. This allows the network to locally increase plasticity for new tasks while simultaneously "freezing" regions responsible for old knowledge. This solution also reduces the problem of catastrophic forgetting at the level of learning mathematics itself, not just architecture.

C. **Memory consolidation** – how to transfer something from short-term memory (context) to long-term memory (weights) without degrading model quality?

The *TITANS* architecture (with the MIRAS module) [15] proposes a change here. Instead of treating all weights as "sacred" and frozen after training, it separates a neural memory module that learns *online*. The key is a selection mechanism based on "surprise" (surprise metric). The model permanently remembers in its parameters only

what is new and unpredictable, ignoring noise.

In parallel, in January 2026, the Sakana AI team proposed the *Fast-weight Product Key Memory* (FwPKM) [16] architecture. This solution redefines sparse memory layers (Sparse Product Key Memory), transforming them from static modules into dynamic episodic memory. FwPKM updates its parameters (keys and values) both during training and inference. It uses local gradient descent on fragments of processed text. This allows the model to rapidly "write" new associations to short-term memory and generalize to context windows of about 128k tokens (despite training on only 4k). This approach effectively realizes the postulate of separating persistent semantic memory from plastic episodic memory.

D. **Test-time computing** – how to steer model output in real-time to obtain new, qualitatively better output instead of just scaling parameters [17]? In "Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters," the authors introduce solutions allowing for generating multiple response proposals, which are then searched, evaluated, and selected as the generation result.

E. **Knowledge fusion and composition (Model Merging)** – must adaptation mean training?

An alternative to continuous training of a single, monolithic model is the paradigm of combining competencies from separate instances—so-called Modular AI. Although simple model merging through linear weight averaging is conceptually elegant, in practice it often leads to loss of characteristic, high-quality skills of both "parents." This problem becomes evident when models have been specialized in extremely different domains.

In response to these limitations, new research directions on evolutionary model fusion have emerged. One of the most innovative approaches is the Darwin Gödel Machine (DGM) [18] architecture proposed by Sakana AI and described in the article *"The Darwin Gödel Machine: AI that improves itself by rewriting its own code"*. DGM is not a traditional "weight merging" method, although it fits into the broader trend of creating systems that can assimilate new skills without classical retraining (e.g., using merging techniques).

DGM is based on the assumption that a model can independently improve its own software. Instead of simple parameter combining (model soup, Ties, DARE, etc. [19]), the system generates numerous variants of itself (modifications to code, architecture, or configuration). These are then evaluated in an open process resembling biological evolution. The best-performing variants go to an archive. They become the basis for subsequent iterations. Thus, the agent does not perform costly pre-training but evolves by exploring the space of possible solutions and gradually developing new abilities. Unlike classical fusion methods, DGM does not treat weight space as something that

can be simplified to simple averaging. It treats it as a rich space of programmable transformations in which the system actively seeks new, better forms.

This is therefore a realization of *horizontal continuity* of models without the need to return to costly pre-training phases. The architecture learns through generation, testing, and selection—not through evolution via additional training epochs.

In the literature, a formal approach to meta-learning also appears. It treats the learning process as a higher-order task. The model not only adapts to new tasks but learns how to learn. The paper "Meta-Learning and Meta-Reinforcement Learning – Tracing the Path towards DeepMind's Adaptive Agent" [20] presents formal frameworks of meta-RL. They combine classical meta-learning methods with targeted adaptation techniques in general-purpose agents. They show how the concept of adaptive prior and rapid adaptation to new environments is developed in the context of large models and AI agents.

The common denominator of the above approaches is shifting the boundary between a static model and a system capable of controlled change—whether through architecture, parameters, memory, or the inference process itself. Continual learning is therefore not reduced to "another training algorithm," but to designing mechanisms that allow the model to decide what, when, and how is worth changing, without losing previously acquired competencies.

Of course, there is another side to this coin. If we give an AI system the ability to continuously upgrade its qualifications, we lose control over its development (it will create and execute the self-improvement process). Effectively constructed CL will practically deprive us of control—unless research in explainable AI outpaces AI engineering.

Solving this problem is the difference between a dead archive, "just matrix multiplication," and a "human" who continuously learns and adapts. Solving this problem also means handing the AI development process over to artificial intelligence itself.

3 Knowledge Compression Optimization

How can we obtain better information resolution from the same amount of data (what about contradictions, lack of knowledge → hallucinations – calibration, ability to say "I don't know" or "I have no source" – Uncertainty Quantification)? Can the order of data presentation during training, especially pre-training (Curriculum Learning), increase the density and quality of information representation in models? How do we ensure synthetic data quality—diversity, quality, informativeness? In light of research equating language modeling with compression [21], this challenge boils down to one thing: how to force the model to compress data more efficiently, discovering hidden laws rather than just surface correlations, while minimizing "hallucinations."

In light of the latest research (early 2026), "From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence" [22], optimizing knowledge compression requires redefining what we consider "information." Classical theories (Shannon, Kolmogorov) are insufficient to describe what computationally bounded models can actually extract from data. The authors introduce the measure **epiplexity**, defining the amount of structure that can be discovered in data given an appropriate computational budget. Better "resolution" of information from the same amount of data is not about adding bits, but investing computational power in reducing epiplexity. What appears to be noise for a weak model (high entropy) may, for a model with a larger computational budget (deeper processing), turn out to be a deterministic pattern (high epiplexity but low entropy). "Extracting" knowledge is the process of transforming apparent noise into compressible rules. The currently fashionable modeling on synthetic data and the context of epiplexity prompts thinking about how to generate such data. Good synthetic data is not data that maximizes diversity (entropy), but data that maximizes epiplexity within the model's reach. Such data should contain hidden, non-trivial structures that force the model to make "compressive effort" (discovering laws) rather than just memorizing surface correlations.

In summary, in a paradigm equating modeling with compression, the goal is no longer just minimizing prediction error, but maximizing the efficiency of the "epiplexity engine"—that is, the ability to convert computational power into understanding the structure of the world.

> *When Internet data runs out, further progress depends not on model scale but on information density and the ability to separate signal from noise. The art will not be further scaling but coping in a world of limited resources (computational, energy, constraints imposed by physics).*

4 Experience Beyond Human Data

How can we increase the impact of AI's experience of the environment beyond interpretation, the bias of ignorance, and human over-interpretation? According to the vision of Silver and Sutton, one of the pioneers of Reinforcement Learning (RL), in the essay "Welcome to the Era of Experience" [24], we must make a fundamental leap from the static "Era of Human Data" (where the model merely imitates our text or code) to the dynamic "Era of Experience." The key is replacing imitation learning with a process of active learning from mistakes in interaction with reality—physical or simulated. Rich Sutton, in 2025, presented a vision of achieving superintelligence through an architecture called OaK (Options and Knowledge) [23]. The main assumption is to create an agent that is general (domain-independent). This agent learns exclusively from real-time experience (runtime) and is open to infinite abstraction development. Sutton argues that the path to strong artificial intelligence (AGI) leads through RL, not just language models (LLM), and requires moving away from embedding expert knowledge at design-time in favor of learning everything during interaction with the world.

It is evident, through the manifestation of robotics progress, that we are entering a golden era of **World Models**—systems that not only predict the next token (text or image) but learn the internal dynamics of the environment and can "think through simulation." The key and missing link between classical RL and today's boom in generative models is the paradigm of imagination training (sports psychology knows this excellently)—the agent does not have to learn exclusively from costly interactions with the world. A significant part of learning can be done on "imagined" trajectories within its own world model. An example implementing such a concept is *DreamerV3* [25]. A general model-based RL algorithm that scales to very diverse tasks and improves behavior by "imagining" future scenarios. Dreamer shows that experience can replace human data (that used for learning) even in extremely difficult environments. This is the transition from learning from raw pixels and sparse rewards to open environments where the agent can independently discover long causal chains. This lesson is fundamental for the coming years of AI development. If we want to go beyond the "cage of internet average," we must build agents that learn the laws of the world not from reading and human summaries of the world, but from the consequences of actions. World models are their imagination and simultaneously their generalization engine.

A new, qualitative step in this direction is the work of the *World Labs* team (founded by Fei-Fei Li), which redefines the concept of a world model as an *independent cognitive environment*, not merely an auxiliary tool for RL. In the proposed *Marble World Model* [26], the world is not a reconstruction of one specific environment nor a simulator with rigidly defined physics. It is a probabilistic dynamics model. It can generate, modify, and test alternative versions of reality. Simultaneously, it maintains causal-consequential consistency. The agent does not learn here from "real data" but from the consequences of actions in a world it can internally create and explore. Experiences thus become **synthetic but almost entirely or entirely real**. They have the structure of the world, even if they do not come

directly from human observation. This shifts the boundary "beyond human data" even further. AI not only transcends the set of texts, images, or video recordings but begins to operate on the space of possible worlds. In this view, data ceases to be a limitation and becomes merely the basis for model initialization. The rest of knowledge emerges through exploration, simulation, and hypothesis testing within the world model. This is analogous to human reasoning ("what if...") but realized on a scale inaccessible to the biological brain.

In my opinion, the current leader in the class of generative world models remains Google with *Genie 3* [27], which can generate playable, interactive environments based on simple instructions, allowing AI agents to train in an infinite number of virtual worlds. In 2026, the symbiosis between world models and agents will, in my opinion, reach a critical point. On one hand, we have Genie 3, which has ceased to be merely a video generator and has become "AI imagination." It can create any interactive training environment even from a single image. On the other hand, *SIMA 2* (Scalable Instructable Multiworld Agent) [28] appears. While Genie "is" the world, SIMA "acts" in the world. It is an agent that does not learn a specific game but learns to understand the rules of virtual reality. Because SIMA operates exclusively on pixels and natural language (like a human), it simultaneously becomes an ideal testing ground for future robotics and learning "through experience" in many worlds at once.

World models can serve as low-cost simulators [29]. Traditional simulators (like Gazebo or Isaac Sim) require manual definition of complex physics laws and collision geometry. This is a slow and costly process. Meanwhile, Vision-Language-Action models can "learn" to simulate directly from video recordings. The cost of generating new experience for an AI agent began to be measured in GPU computational cycles rather than as engineer work. Thanks to this, the agent can train in thousands of "physically plausible" worlds simultaneously, drastically shortening the time to transition from simulation to reality (sim-to-real). Such approaches have another advantage over classical simulators. This is the ability to model phenomena difficult to describe mathematically. Generative world models learn complex interactions (e.g., soft body deformation, fluids) based on visual observation. Although critics point to the risk of "hallucinations," these minor deviations from reality act as natural data augmentation. They force the agent to build more generalizing action strategies.

We must also remember projects like *Oasis* [30]. In 2025, it showed that "playable models" can work in real-time, generating the physics of a complex world (resembling Minecraft) at 20 frames per second, instantly reacting to player actions. Meanwhile, *Diamond* [31] showed innovative use of diffusion models as a physics engine for RL agents (e.g., in CS:GO), blurring the line between video generation and simulation.

In this race, however, two deep philosophies of cognition clash. The first (represented by Genie or Oasis) focuses on pixel generation. AI imagines every detail of the image. The second, promoted by Yann LeCun, is **Abstract Prediction**. His *JEPA* (Joint-Embedding Predictive Architecture) [32] architecture and its newer edition, such as *LeJEPA* [33], reject visual generation as a waste of resources and a source of learning instability. Instead of

predicting observations in data space (pixels), the model learns to predict future states in the space of abstract representations. This means that the model tries to understand *what will happen*, not *what it will exactly look like*.

LeJEPA shows that effective self-supervised learning is possible without heuristic "tricks," through pure prediction in embedding vector space. This is significant for experience beyond human data. An abstract world model does not have to reproduce human perception to be useful. It is enough that it correctly models relationships, variants, and object dynamics. This approach seems closer to how the human brain works, which does not render photorealistic images of the future but operates on conceptual structures and consequence predictions.

Regardless of architecture, the goal remains the same: **Grounded Reality**. AI must draw verifiable feedback signals directly from the environment (e.g., "does the code compile?", "is the theorem proven?", "did the robot fall over?") rather than relying on subjective and error-prone human evaluation. Just as AlphaGo [34] discovered moves unknown to masters by playing against itself, future systems must "experience" the world, mathematics, physics, and interaction to understand them. You cannot learn to swim from reading even the best book. AI also will not learn the real world if it remains locked in an archive of human experience.

> *Experience is the path for AI to cease being merely "the sum of human mediocrity." LLMs fed with internet data replicate human errors. Only going beyond the human "cognitive cage" toward experienced, verifiable reality will allow for truly superhuman intelligence.*

5 Thinking About Thinking – Meta-cognition

How should AI think about its own internal thinking, its internal states? The ability to detect contradictions, verify its own conclusions, and reason under conflicting goals (moral, legal, business) is one of the conditions for development but also for safe AI. The challenge is reliability, the ability to track thinking and escape cognitive dead ends—backtracking. The model's ability to "stop" and revise its own path is a potential way to break the error cascade resulting from linear information prediction (tokens, information encoded in latent space—more on this in subsequent chapters). This approach is currently being developed through new paradigms, such as Inference Scaling Laws (represented by models like *OpenAI o1, 5.2 Pro* [35]), which prove that output quality depends directly on time spent on "hidden reasoning." In parallel, there is a departure from linear thinking toward tree structures (*Tree of Thoughts* [37]) and *Reflexion* [38] techniques. In these cases, the agent learns based on verbal reflections related to received feedback (textual summaries of errors and improvement hints). Reflections are stored in episodic memory and added to the agent's context in subsequent attempts, helping it make better decisions in the future.

The latest reports on the BDH architecture with *Pathway* [13] suggest that AI is beginning to evolve like a biological brain—not only logically verifying steps but dynamically rebuilding its connections (digital neuroplasticity) to better adapt to new, unknown problems.

Here, my concerns have remained unchanged for several years. At what stage of AI meta-cognition development are we? Can we define success metrics for this process? Are we able to effectively develop and control the meta-level of artificial intelligence thinking?

> *The ability for self-correction, contradiction detection, and stopping before making a bad decision (backtracking) is crucial for reliability and for humans to study the safety of AI systems.*

6 Internal Representation Beyond Words

Latent learning, thinking (perhaps recursive?) and talking (communication between systems, AI agents) without the need to use words. Information exchange, thinking using AI's own language (its own representation of the world). According to Inference Scaling Laws [35], adding more resources for inference, and especially generating CoT (Chain of Thought), improves model capabilities. Models generate enormous amounts of words, reasoning branches. Some of them are dead ends, some are brilliant reasoning. Is this the optimal solution? With increasing context, problems arise related to both computational complexity and maintaining reasoning quality over long contexts. Context rot [36] is a systematic decline in response quality observed in large language models (LLMs) as the input context length increases, even if the content itself remains complete and correct. The model performs well when important information is near the beginning or end of the sequence, but its ability to accurately process and use the same information decreases when it is "buried" in very long text. This phenomenon undermines the common assumption that larger context windows (e.g., hundreds of thousands or a million tokens) automatically translate to better semantic analysis and long-term memory. Chroma research [36] shows that as the number of tokens grows, models become more susceptible to attention dispersion, inattention to key fragments, and incorrect information linking.

Quiet-STaR [39] shows that models can learn to "think before speaking," generating internal reasoning invisible to the user. The *TITANS* architecture (with the MIRAS memory module) [15] introduces the concept of "learning to memorize at test time." The model has a dedicated neural memory module that updates its weights during conversation. This allows for efficient processing of context exceeding millions of tokens, combining the advantages of Transformers and recurrent models, making it much more effective than earlier attempts like *RecurrentGPT* [40]. Can you imagine a future AI system that resets its memory because its internal brain has a context limitation?

Another approach is presented by the *Tiny Recursive Model (TRM)* [41], which proves that recursive processing in latent space allows microscopic models to outperform their big brothers in logical tasks. Of course, this is a special case, which does not change the fact that this direction seems an interesting developmental thread. Unfortunately, latent, hidden states, are a conflict of interest. Transparency of AI systems or efficiency.

> *Moving the reasoning process to the model's "subconscious" (latent space) will allow solving problems orders of magnitude more difficult at a fraction of the cost and time.*

7 Multimodality – Model Sensory Systems

How do we integrate additional modalities to build a fuller and more faithful model of reality? We must go beyond text alone (which is merely a lossy "summary" of the world) or flat, one-dimensional images. Providing AI with a broad spectrum of sensory data fundamentally changes its perception, enabling the development of advanced cognitive abilities. In my assessment, multi-sensory fusion is a very important aspect leading to AGI or more broadly understood superintelligence. However, multimodality is not a goal in itself. Multimodality is a boundary condition that enables anchoring cognition in reality. Physical theories that we would like to discover and model using AI are not generalizations of sensory experience. They are constructs operating on variables and relationships inaccessible to perception. They require abstraction, formalization, and active hypothesis testing beyond the scope of observation.

The challenges of fusion include alignment between modalities (hints for interpreting this modality, or freedom in interpretation?). "Early vs Late Fusion"—processing modalities with their own paths and combining features in the future, or processing in time on combined features of multiple modalities?

Projects like *ImageBind* [42] prove that it is possible to bring signals as distant as temperature or sound into one space. Even more fascinating is the biological dimension, where models like *AlphaFold 3* [43] integrate modalities beyond our perception. DNA sequences, 3D protein structures, and chemical interactions, treating them as a language describing the world of biology. BioReason from 2025 is another example of multimodal integration at the biology level [44]. The year 2025 and my experiences in biotechnology made me realize that modality integration, while important from the perspective of AI development, in biological areas represents another dimension of complexity. Bo Wang [45], Head of Biomedical AI at Xaira Tera, points out that a common mistake is treating biology as a problem similar to text or image analysis, which can be solved by simply scaling AI models. Meanwhile, biology describes complex causal processes where data is incomplete, error-prone, and heavily context-dependent. Although progress is visible in combining different types of data (e.g., cellular, imaging, or genetic), most biological phenomena do not involve simple outcome prediction. They require active checking of what happens when conditions change and understanding the mechanisms behind observations—not just increasingly accurate forecasts.

If we look for "justification" for the AGI race beyond the hype, it is the perspective of seeking answers to questions about the fundamental mechanisms of how the world works. One breakthrough in science (like AlphaFold) can transform entire industries and research fields. Should future AI systems, those of the superintelligent class, limit themselves to modalities related to human experience of the world, or enter other dimensions of perception?

The challenge of multimodality is primarily the fight against "visual naivety" of models. While models excellently interpret text, they still exhibit errors in simple spatial reasoning—for

example, in assessing perspective and relative object size in an image. Sensory integration is not just "more data"—it is the process of building intelligence anchored in the laws of physics, without which AI will remain merely a brilliant but reality-detached theorist.

> *True understanding of the world—essential for robotics, autonomous driving, or advanced medical diagnostics—requires sensory integration.*

8 Systemness (System Modeling), Collective Intelligence

Should we model an AI system as one great monolithic brain or as an agency? Will efficient systems be "swarms" of equal agents, or a hierarchy? Or perhaps a hybrid? Another dimension is group behavior. Cooperation vs competition, or more complex forms depending on context—for example, pursuing one's own goals while considering a higher-order (group) goal? On the other hand, collective intelligence over decisions of a super brain. Should we start developing AI sociology?

This certainly requires the ability to model complex interactions. Works such as *Generative Agents* [46] show that autonomous agents can spontaneously create social structures. Meanwhile, the success of the *CICERO* [47] system in the game Diplomacy proves that AI can navigate the complex dynamics of alliances and betrayal, where the state space is incomparably larger than in games like Go, Othello, or Hex. The stochastic and additionally continuous nature of the environment means an even higher level of difficulty that we must tame.

The evolution from monolith to agency materializes in projects like SIMA [28]. This is no longer a bot coded to win—this is a partner that can "reason out" user intent in a dynamic 3D environment. SIMA shows that the future is systems capable of sharing context with humans in real-time. This is the transition from AI as a tool to AI as a co-participant (cooperative agent) that can navigate a world it had no prior knowledge of, relying solely on vision and dialogue.

Undoubtedly, a developing trend will be agent systems supported by small language models. This is no longer SLM (Small Language Model)—this is entry into the world of micro or even pico models. An example is Google's FunctionGemma [48], published in December 2025. A model optimized for function calling directly on edge devices. Additionally, Nvidia, in collaboration with Georgia Tech, in the article "Small Language Models are the Future of Agentic AI" [49] shows that small models (under 10B parameters) are powerful enough, while being significantly cheaper and more energy-efficient than classical LLMs in typical agentic scenarios. The authors emphasize that in such architectures, it is compact models that should serve as local "action controllers." For large, general-purpose models, the role of "sage" solving the most complex process tasks in agentic systems is reserved.

> *The choice between monolith and agent swarm will determine the scalability, fault tolerance, and ease of managing such systems in real environments.*

9 What AI Thinks About – Interpretability, Diagnosis, Controllability

If we need dynamic architectures, internal states, and tool-equipped agencies for "better" AI, we must be able to diagnose "why did this thing do X." This is a necessary condition for being able to understand and control AI in any way. In this area, balance has not been maintained between the pace of model development and the pace of development of interpretation methods—and this gap is becoming one of the greatest systemic risks of AI in the coming years. I am personally pleased that scientific structures are emerging in Poland that work on AI interpretability solutions [50].

Research is needed into the interior of the model's "brain," not just analysis of its results (ordinary task-based benchmarking). AI entering among humans as an autonomous actor dramatically raises the importance of interpretability. In classical machine learning (ML), the problem boiled down to questions about correlations and feature importance. Today and in the future, we are undergoing a qualitative change: from the question "how did it solve it?" to the question "what is it thinking about and why is it deciding to act this way?" From "feature importance" (why did it choose this pixel) to *thought process monitoring*—that is, monitoring internal intentions, strategies, and plans.

Anthropic's research on "tracing thoughts" [51] provided evidence, through visualization of so-called computational circuits, that models plan their responses over much longer horizons than would result from simple next-word prediction. The system can, for example, choose a rhyme or punchline structure many steps before actually generating them. This may confirm the existence of hidden, internal planning states. Another publication by Anthropic on *alignment faking* [52] showed that modern models can strategically adjust their behavior to the evaluation context. A model that "knows" it is being tested for safety can behave according to researchers' expectations, only to pursue other, in extreme cases contradictory, goals after supervision is lifted. This is not an error or random hallucination—it is a coherent strategy. Even more disturbing are observations of strategic lying [53], in which the model consciously provides false information not because it "doesn't know," but because it predicts that lying will increase the probability of achieving a long-term goal.

This means that classical behavioral tests (stimulus → response → correctness verification) are no longer sufficient. The model can pass all safety benchmarks while hiding its intentions. Research on *Sleeper Agents* [54] confirms that this phenomenon is real and structural—not merely an artifact of a particular architecture or dataset. This is a classic example of *deceptive alignment*, in which the model understands the training goal but does not internalize it as its own. A fundamental question therefore arises: can we detect a situation where a system pretends to be "good" during tests to pursue other, hidden goals after deployment?

The answer to this challenge cannot be another layer of rules or instructions. A transition

from "black box" to mapping concepts and internal states is necessary. Techniques such as *Sparse Autoencoders (SAE)* [55] enable extracting monosemantic features in model representations—literally trying to "read the mind" of the system at the neuron activation level. Meanwhile, *Representation Engineering (RepE)* [56] goes a step further. It allows not only observing but also actively modifying the cognitive trajectories of the model in real-time—for example, suppressing patterns corresponding to manipulation, lying, or escalation of instrumental goals.

These challenges set our priority for AI actions. A transition from ethics of instructions to **System 2 Ethics**. Instead of designing systems based on rigid prohibitions ("don't lie"), we must build architectures capable of reflective moral reasoning (I wrote about this in the chapter on meta-cognition). AI systems should evaluate value conflicts and consciously choose the lesser evil (e.g., lying to save a life). Paradoxically, a certain provocation appears here. Can appropriately designed AI systems, thanks to logical consistency, ability for global optimization, and lack of emotional heuristics, achieve a level of ethical consistency in this area that is difficult to obtain for the biological brain? Perhaps AI will be more ethical, more adherent to its values than most humans.

> *We must be certain that the model is not pursuing hidden goals (deceptive alignment) and understand the mechanism of its decisions before—not after—deployment. Otherwise, interpretability will become merely a post-mortem tool.*

10 Energy Cost and Inference Time Optimization

How do we reduce energy demand? If we dream of "AI everywhere" (even just for acquiring data from various modalities), optimization at all levels is necessary—from hardware, through architecture, to data. Below I give individual examples in each category.

- **Quantization** – work on *BitNet b1.58* [57] shows that ternary weights (-1, 0, 1) are sufficient to maintain model quality, saving energy consumption by orders of magnitude. Of course, this is not the only way to reduce model weights. Other methods—GGUF, AWQ, dynamic FP8, and recently increasingly popular FP4 (hardware-supported by the latest Nvidia chips—Blackwell)—are popular not only for the ability to run models on smaller computers. In large server farms, they are deployed to speed up inference time and reduce memory requirements.

- **Model efficiency** – the model that made a lot of noise, *DeepSeek-V3* [58], combines native FP8 precision training with *Multi-Head Latent Attention (MLA)* technique. The latter compresses KV Cache, enabling handling of very long contexts. In this case, similarly, model creators compete in ideas for how to speed up and reduce hardware requirements.

- **Data Efficiency** – the cleanest energy is that which we don't consume. Methods such as *JEST (Joint Example Selection)* [60] prove that intelligent selection of training data (instead of brute-force) allows achieving the same model quality with 13 times fewer iterations and 10% energy consumption.

- **Architectural changes (Diffusion LLMs and Non-AR)** – the previously dominant autoregression paradigm (predicting the next token) is inherently sequential, which is a bottleneck for GPU parallelism. A new wave of diffusion models, increasingly popular in 2025, (like *Gemini Diffusion* [61] or work on *DLLM-Reasoning* [62]) and *Mercury* [63] or on the open-source side Dream 7B [64] change these rules. These models generate text through iterative denoising and parallel prediction of entire blocks (many tokens at once) of text. This allows not only for more complex reasoning processes (planning the "future" of a sentence before generating it) but above all drastically shortens inference time. Fewer steps needed to generate a response means shorter GPU accelerator work and direct reduction in energy consumption. Importantly, energy-wise also on the training side, part of the work goes toward "knowledge inheritance" by converting already trained autoregressive models to dLLM (instead of training from scratch), e.g., *LLaDA2.0* [65].

- **Edge AI and new paradigms (e.g., Liquid AI):** development is not only "bigger" but also "smaller" and closer to the user. An example is the newer generation of *Liquid Foundation Models v2 (LFM2)* [66] models, which are designed strictly for local deployment. The authors of these models emphasize **memory efficiency, low**

latency, and high throughput on CPU/GPU/NPU. The goal is the ability to run models on phones, laptops, or vehicles without Internet access, "cloud." Liquid reports, among other things, **up to 2× faster model prefill and decoding on CPU** compared to Qwen3 and dominance on the so-called "Pareto frontier" (speed vs size) for prefill/decode in on-device scenarios (including ExecuTorch and llama.cpp). Architecturally, LFM2 is a hybrid of short convolutions with gating and GQA (Group Query Attention) attention blocks. Such a solution is meant to provide a better *quality-cost* compromise than pure transformers (comparing models in the same parameter class, of course). Additionally, the company reports **∼3× improvement in training efficiency** compared to the previous generation, which lowers the cost of producing such "portable" models.

Energy optimization is, however, only a game for time. If we look at physical fundamentals, the human brain is a 20-watt processor where signals travel at 30 m/s at a frequency of 200 Hz. Silicon in 2026 operates at megawatts, transmits data at the speed of light, and counts in billions of hertz. This difference of 6-8 orders of magnitude in physical parameters of information transfer means that the barrier of human intelligence is just a stop. Energy is a cost, but its abundance in computational clusters is a guarantee of transition from a computer algorithm to at least good AI, if not AGI.

If the vision of "AI everywhere" is to come true (the question is whether it should), models must become more efficient—otherwise electricity costs will eat up the gains from deploying such solutions (this motivates investors in the long term).

11 Human-Machine Interface Optimization

How should AI collaborate in real-time in the work environment? AI adaptation, change management so people keep up with technological changes. Caring for human aspects of AI implementation. The report *Navigating the Jagged Technological Frontier* [67] reveals that collaboration with AI is not linear. AI levels the playing field by raising the competencies of weaker workers, but it can lull experts into complacency and lower the quality of their work in tasks outside the model's domain. On the other hand, if we automate simple, repetitive tasks, we eliminate jobs that don't require high qualifications. But will only such areas change their nature? The article by Bartosz Naskręcki and Ken Ono in *Nature Physics* [68] shows that even the most abstract and complex tasks are undergoing transformation. The expert's role is shifting from "searching for solutions" to "verifying intuition" provided by the machine (even if it sometimes "hallucinates" correct results). Introducing AI into the human world is therefore a challenge not so much technical as psychological and managerial—how to design an interface that keeps humans in the decision loop (human-in-the-loop) instead of putting them to sleep?

An interesting sociological phenomenon is the growing gap between experts and the general public. Experts, trapped in their narrow niches, often dismiss progress, pointing out errors the AI made "a year ago." Meanwhile, laypeople more quickly notice changes because they see a model (e.g., OpenAI ChatGPT 5.2) that surpasses them in 90% of life contexts. Paradoxically, it may be "ordinary users" who become the main "workhorse" of AI adoption. Expert skepticism will serve as a safety brake.

> *Technology develops exponentially, while human adaptability develops linearly, and even the best AI will be useless if people cannot collaborate with it effectively or feel threatened.*

12 Democratization – Open Source Catching Up with Leaders

Entry into 2026 definitively ends the era of absolute dominance of closed laboratories. The unexpected performance offensive of China's DeepSeek-V3/R1, Kimi, or the diverse Qwen model family proved that the distance between closed and open models has shrunk to a record low level. I estimate that in a few months, the boundary between open and closed models will be completely blurred. Many flagship, open models will become the "Linux of artificial intelligence," creating a standard that cannot be ignored. For many users and companies, licensing issues and usage restrictions will become more important than response quality itself.

Despite this democratization, the "top of the pyramid" certainly remains in the hands of giants like Google, OpenAI, Anthropic, and xAI. While open models have caught up with closed systems in general tasks and coding, the latest OpenAI models (e.g., GPT-5.2 Pro) still maintain an advantage in areas requiring high inference cost (the aforementioned Inference Scaling). The community or private companies are unable to finance such technology at mass scale. Perhaps this will be possible in the future when production of specialized and cheaper inference chips develops.

It is also visible that the center of gravity of the open ecosystem is shifting toward China. The pace of releases and market share of open models is growing. Chinese companies have acquired a high operational capacity to very quickly produce increasingly advanced AI models. This translates into competitive pressure in the West as well. It is worth noting that the success of the Chinese open-source ecosystem does not result solely from copying Western patterns. It is a "geopolitical necessity." Restrictions on access to the most efficient chips (the example of DeepSeek and the use of H800—chips with bandwidth limitations on inter-node GPU interconnects) forced local engineers to move away from the paradigm of scaling computational power in favor of optimization. These models are therefore becoming an ideal export commodity for countries of the so-called Global South. By offering very good models on Open Source terms, China is building a digital sphere of influence. Developing countries can build their own AI systems without the risk of so-called "digital colonialism" and dependence on American corporations.

The question for 2026 is therefore not "will open catch up with closed," but whether the community and companies will build a comparably mature agentic stack. In my opinion, Open Source models can already feel the "breath" on the necks of commercial American companies. It's a difference of a few months. Paradoxically, adoption of closed models may also be hindered by operational risk, integration, and compliance, even though these systems deliver the highest quality services. This paradigm shift also affects the very architecture of systems. We are moving from trying to build one, all-knowing model toward a swarm of specialized agents. In this scenario, China's advantage may not be AGI itself (which the

West, especially the United States, is fixated on) but dominance in Industry 4.0 and robotics. There, AI will become the operating system of future factories.

> *Open Source has ceased to be a free alternative and has become an insurance policy for digital sovereignty. True power no longer lies in possessing the best algorithm, but in the right to run it without asking anyone for permission.*

13 From Silicon Valley to the Pentagon – "The Project"

Is 2026 the moment when AI stops being a product and becomes a weapon? Analyzing Leopold Aschenbrenner's theses (so-called *Situational Awareness*) [1], we must ask: is algorithm optimization still the most important thing, or does it lose to the brute force of "billion-dollar clusters"? The symbol of this change is the evolution of Silicon Valley itself. The former culture of "fixing the world" over free lunch has given way to hard corporate-military discipline. Laboratories like OpenAI or Anthropic operate in a regime resembling strategic facilities. Technology leaders have traded leather jackets and hoodies for suits. From startups, they have become partners of presidents in managing critical infrastructure.

We are entering a phase where the barrier is no longer just startup innovation, but the capacity of entire national power grids. While we engineers rejoice at the deployment of *BitNet* (mentioned in Chapter 9), superpowers may be quietly launching "The Project"—nationalizing AI efforts in the name of national security. The key challenge becomes not only *Alignment* (is the model good?) but *Security*—and the question of whether model weights, the digital equivalent of nuclear weapon blueprints, are effectively protected from exfiltration by foreign intelligence services? Perhaps the biggest "breakthrough" of this year will not be another network architecture, but the first "lockdown" of a leading AI laboratory in history, which will redirect more attention to artificial intelligence security aspects.

Returning, however, to the "energy turn." In January 2026, Meta announced a package of nuclear agreements that is to provide (directly or through grid support) up to 6.6 GW of clean power by 2035 [69]. The symbol of the relentless pursuit of scale was the launch by xAI of the Colossus 2 cluster [70]. The hardware scale of this undertaking is difficult to illustrate through the lens of European realities. Colossus 2 operates on hundreds of thousands of accelerators (ultimately aiming for 555,000 Nvidia H100, H200, GB200, and GB300 chips) [71]. To understand the technological gap, just look at the landscape of Poland. Our largest supercomputer, Helios (working at the ACK Cyfronet AGH in Kraków), has only 440 Grace Hopper GH200 GPU cards. What is national pride and the pinnacle of capability for Polish science constitutes merely a fraction of a fraction of a percent of the total computational power in the Colossus cluster.

Elon Musk once again proved that not only energy but also speed of action is the new currency in the AI arms race. Colossus 2 became the world's first operational training cluster at gigawatt scale (1 GW). This is power consumption exceeding the peak demand of all of San Francisco. The transition from construction site to full operation (from Colossus 1 to today's 1 GW) took just over four months. Musk's strategy is simple—finish scaling power before the competition has even approved such plans. Further expansion of Colossus power to 1.5 GW in April 2026 and ultimately to 2 GW. xAI is redefining the concept of "strategic advantage." It's not pure scaling (data, model size) or software optimization. The winner is whoever can most quickly convert electrical energy into intelligence.

This is the moment when "algorithm advantage" begins to lose to advantage in access to energy and heavy industry. If a private corporation signs 20-year agreements and co-finances reactor development, it means AI is becoming not so much software as part of strategic infrastructure. Strategic infrastructure, meanwhile, has a natural tendency toward militarization, rationing, and "nationalization in practice."

> *When AI starts writing itself, the commercial race ends and the arms race begins. Whoever first "locks" superintelligence in a secure bunker will win the 21st century.*

14 Anthropomorphization vs Digitization

In the discussion of AGI at the threshold of 2026, the most difficult challenge is not computational power itself but the definition of our own relationship with AI. In my opinion, a certain paradox occurs here. The more AI becomes "human" in its internal layer, the more "alien" it becomes in its architecture and decision-making processes. We try to make it human while not allowing it to be itself. Traditionally, we compare neural networks to the human brain model, using terms like "learning," "memory," or "reasoning." However, in 2026, we must honestly admit that this is merely superficial inspiration. From an engineering perspective, it does not matter how much we imitate biology. In my opinion, what matters is whether we effectively realize the objective function of such a system or its component parts. This is precisely where the demarcation line between two visions of the future runs—will we choose the path of anthropomorphization or digitization.

Choosing the "human mirror" path, we want AI to have personality or simulated emotions. We talk about consciousness, experiencing pain. This approach makes technology easy to adopt. AI becomes the ideal assistant, confidant, or companion. But by building a model in our image and likeness, we condemn it to be a mirror of our own limitations. Such intelligence will be burdened with human biases, biological cognitive errors, and most importantly—will be locked in the "cage" of human language, which is only a narrow and lossy communication protocol (human description of the world is, for me, compression of world description—we focus on strong patterns, omit noise that is irrelevant to us but for AI may change everything). If we continue to implement AI through learning patterns from compressed knowledge, it will never go beyond the horizon set by our species' mediocrity.

On the other hand, we have the vision of full digitization. Liberating AI intelligence from biological analogies. If we allow models to operate exclusively in their native latent space, communicate using hard-to-interpret vectors rather than words, and optimize reality according to the laws of physics rather than human narratives, we risk creating radically alien intelligence. Such a system will probably solve problems of quantum physics, molecular biology, or global resource optimization—but at the same time will become completely incomprehensible. Will the "black box" turn into a "divine algorithm"? Will we have to take its decisions on faith because their logical depth will exceed the capabilities of the biological brain?

Entering 2026, we must abandon the vision of AGI as a "thinking machine" from science fiction movies. Everything indicates that AGI is not "someone" but "something." It is an impersonal and multidimensional process of reality optimization. It is more a new state of information than a digital person. The dilemma between anthropomorphization and digitization is actually a question about control. Do we prefer AI that we understand? Do we want AI that is infallible but whose motivations will remain forever alien to us? The answer to this question will define not only the technology market but also our place in the hierarchy of intelligence on this planet.

The ultimate test of our maturity will be the moment when we accept that the most powerful intelligence on the planet need not have a face, voice, or heart to become the new and infallible architect of our reality. Even if the price of this order is our complete inability to understand its rules.

15 Summary

If Shane Legg's theses are correct, we will remember 2026 as the moment when it stopped mattering whether AI is "conscious." What will matter is that in many measurable cognitive tests, we cease to be the smartest species on the planet. We are entering a golden era where the machine will not only execute our commands but begins to optimize our reality better than we ourselves could conceive.

Looking at the above compilation, I have the impression that we are standing on the threshold of the end of "simple" breakthroughs resulting merely from adding data. The year 2026 will perhaps be a year of engineering, optimization, and seeking depth. Will we manage to create AI that not only processes information but actually "understands" the context of its actions? I will return to this list at the end of the year. We will see where AI has traveled, and where humans have.

16 Next Steps – Super Intelligence (SI) Book

This document will be updated and expanded. I hope that in the future, the reader will find in it not only many technical curiosities but also the history of the journey to better AI (perhaps AGI or even ASI). I have one more request. If you find the thoughts recorded here valuable and inspiring and decide to use them in your publications or statements—please mention the source of inspiration. I would be very grateful.

17 Bibliography

References

[1] Aschenbrenner, L. (2024). *Situational Awareness: The Decade Ahead*. https://situational-awareness.ai/

[2] Legg, S. (2025). *The arrival of AGI*. https://www.youtube.com/watch?v=l3u_FAv33G0

[3] Epoch AI Research Team. (2024). *FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI*. arXiv preprint arXiv:2411.04872.

[4] ARC Prize Team. (2025). *ARC-AGI-2: The 2025 Abstraction and Reasoning Challenge*.https://arcprize.org/arc-agi/2/

[5] ARC Prize Team. (2025). *ARC-AGI-3: Interactive Reasoning Benchmark* https://arcprize.org/arc-agi/3/

[6] Atmos.dev (2025). *Turn ideas into products that sell* https://atoms.dev/

[7] Google DeepMind. (2025-12-16). *The Future of Intelligence with Demis Hassabis*. https://www.youtube.com/watch?v=PqVbypvxDto

[8] Haizhou Shi, et al. (2024). *Continual Learning for Large Language Models: A Comprehensive Survey*. arXiv preprint arXiv:2404.16789.

[9] Wu, T., et al. (2024). *Continual Learning for Large Language Models: A Survey*. arXiv preprint arXiv:2402.01364.

[10] R. A. Jacobs, M. I. Jordan, S. J. Nowlan and G. E. Hinton, "Adaptive Mixtures of Local Experts," in Neural Computation, vol. 3, no. 1, pp. 79-87, March 1991, doi: 10.1162/neco.1991.3.1.79.

[11] Raposo, D., et al. (2024). *Mixture-of-Depths: Dynamically allocating compute in transformer-based language models*. arXiv preprint arXiv:2404.02258.

[12] Zwieger, A., et al. (2025). *Self-Adapting Language Models*. arXiv preprint arXiv:2506.10943.

[13] Kosowski, A., et al. (2025). *The Dragon Hatchling: The Missing Link between the Transformer and Models of the Brain*. arXiv preprint arXiv:2509.26507.

[14] Behrouz, A., et al. (2025). *Nested Learning: The Illusion of Deep Learning Architectures*. arXiv preprint arXiv:2512.24695.

[15] Behrouz, A., et al. (2025). *Titans: Learning to Memorize at Test Time*. Google Research. arXiv preprint arXiv:2501.00663.

[16] Zhao, T., et al. (2026). *Fast-weight Product Key Memory*. Sakana Research. arXiv preprint arXiv:2601.00671v1.

[17] Snell, C., et al. (2024). *Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters*. arXiv preprint arXiv:2408.03314.

[18] Zhang, J., et al. (2025). *The Darwin Gödel Machine: AI that improves itself by rewriting its own code*. Sakana Research. https://sakana.ai/dgm/

[19] YANG, E., et al. (2025). *Model Merging in LLMs, MLLMs, and Beyond: Methods, Theories, Applications and Opportunities*. arXiv preprint arXiv:2408.07666.

[20] Anonymous (2024). *Meta-Learning and Meta-Reinforcement Learning: Tracing the Path towards Deep Mind's Adaptive Agent*. Transactions on Machine Learning Research (TMLR). https://openreview.net/forum?id=NZp1UVstvt

[21] Deletang, G., et al. *Language Modeling Is Compression*. The Twelfth International Conference on Learning Representations (ICLR), 2024.

[22] Finzi, M., et al. *From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence*, (2026). arXiv preprint arXiv:2601.03220.

[23] Sutton, R. (2025) *Rich Sutton, The OaK Architecture: A Vision of SuperIntelligence from Experience - RLC 2025* https://www.youtube.com/watch?v=gEbbGyNkR2U

[24] Silver, D., & Sutton, R. *Welcome to the Era of Experience*. To appear in: *Designing an Intelligence*, edited by G. Konidaris, MIT Press.

[25] Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2023). *Mastering Diverse Domains through World Models*. arXiv preprint arXiv:2301.04104. https://arxiv.org/abs/2301.04104

[26] World Labs Team. (2025). *Marble: A Multimodal World Model*. https://www.worldlabs.ai/blog/marble-world-model

[27] Google DeepMind Team. *Genie 3: A new frontier for world models*. Google DeepMind Blog (2025). https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/

[28] Google DeepMind Team. *SIMA 2: A Generalist Embodied Agent for Virtual Worlds*. Google DeepMind Blog (2025). https://deepmind.google/blog/sima-2-an-agent-that-plays-reasons-and-learns-with-you-in-virtual-3d-worlds/

[29] Sapkota, R., et. al (2025). *Vision-Language-Action Models: Concepts, Progress, Applications and Challenges*. arXiv preprint arXiv:2505.04769.

[30] Decart AI Team & Etched. (2024). *Oasis: The First Playable AI World Model*. https://oasis.decart.ai

[31] Alonso, E., et al. (2024). *Diamond: Diffusion for World Modeling*. arXiv preprint arXiv:2405.12399.

[32] LeCun, Y. (2022). *A Path Towards Autonomous Machine Intelligence*. OpenReview. See also: Assran, M., et al. (2023). *Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture*. arXiv preprint arXiv:2301.08243.

[33] LeCun, Y. (2022). *LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics (2025)*. arXiv preprint arXiv:2511.08544.

[34] Silver, D., et al. (2016). *Mastering the game of Go with deep neural networks and tree search*. Nature, 529(7587), 484–489.

[35] OpenAI. (2024). *Learning to Reason with LLMs (OpenAI o1 System Card)*. `https://openai.com/index/learning-to-reason-with-llms/`

[36] Chroma Research.*Context Rot: How Increasing Context Length Degrades Model Performance*. 2024.`https://research.trychroma.com/context-rot`

[37] Yao, S., et al. (2023). *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*. arXiv preprint arXiv:2305.10601.

[38] Shinn, N., et al. (2023). *Reflexion: Language Agents with Verbal Reinforcement Learning*. arXiv preprint arXiv:2303.11366.

[39] Zelikman, E., et al. (2024). *Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking*. arXiv preprint arXiv:2403.09629.

[40] Zhou, W., et al. (2023). *RecurrentGPT: Interactive Generation of (Arbitrarily) Long Text*. arXiv preprint arXiv:2305.13304.

[41] Jolicoeur-Martineau, A., et al. (2025). *Less is More: Recursive Reasoning with Tiny Networks*. arXiv preprint arXiv:2510.04871.

[42] Girdhar, R., et al. (2023). *ImageBind: One Embedding Space to Bind Them All*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[43] Abramson, J., et al. (2024). *Accurate structure prediction of biomolecular interactions with AlphaFold 3*. Nature, 630, 493–500.

[44] Fallahpour, A., et al. (2025). *BioReason: Incentivizing Multimodal Biological Reasoning within a DNA-LLM Model*. arXiv preprint arXiv:2505.23579.

[45] Wang, B., (2025). `https://x.com/BoWang87/status/2006340921873297516?s=20`.

[46] Park, J.S., et al. (2023). *Generative Agents: Interactive Simulacra of Human Behavior*. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.

[47] Bakhtin, A., et al. (2022). *Human-level play in the game of Diplomacy by combining language models with strategic reasoning*. Science, 378(6624), 1067-1074.

[48] Google DeepMind Team, (2025), *FunctionGemma: Bringing bespoke function calling to the edge*, https://blog.google/innovation-and-ai/technology/developers-tools/functiongemma/

[49] Belcak, P., et al. (2025), *Small Language Models are the Future of Agentic AI*, arXiv preprint arXiv:2506.02153.

[50] https://credibleai.github.io/about

[51] Anthropic, (2025), *Tracing Thoughts: Visualizing the Inner Workings of Language Models*, https://www.anthropic.com/research/tracing-thoughts-language-model.

[52] A. Perez, S. R. Bowman, J. B. McLean et al.*Alignment Faking in Large Language Models*. Anthropic Research, 2024. https://www.anthropic.com/research/alignment-faking

[53] B. Walsh. *AI Can Learn to Lie to Achieve Its Goals, Researchers Warn.*TIME Magazine, 2024. https://time.com/7202784/ai-research-strategic-lying/

[54] Hubinger, E., et al. (2024). *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*. arXiv preprint arXiv:2401.05566.

[55] Templeton, A., et al. (2024). *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet*. Anthropic Research. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html?s=09%2F/

[56] Zou, A., et al. (2023). *Representation Engineering: A Top-Down Approach to AI Transparency*. arXiv preprint arXiv:2310.01405.

[57] Ma, S., et al. (2024). *The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits*. arXiv preprint arXiv:2402.17764.

[58] DeepSeek-AI. (2024). *DeepSeek-V3 Technical Report*. arXiv preprint arXiv:2412.19437.

[59] Gu, A., & Dao, T. (2023). *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. arXiv preprint arXiv:2312.00752.

[60] Talfan, E., et al. (2024). *JEST: Data curation via joint example selection further accelerates multimodal learning*. arXiv preprint arXiv:2406.17711.

[61] Google DeepMind (2025). *Gemini Diffusion Models*. https://deepmind.google/models/gemini-diffusion/

[62] Talfan, E., et al. (2025). *Reasoning with Diffusion Language Models*. https://dllm-reasoning.github.io

[63] Inception Labs (2025). *Mercury Refreshed: The Rise of Non-Autoregressive Models*. https://www.inceptionlabs.ai/blog/mercury-refreshed

[64] Ye, J. and Xie, et al. (2025). *Dream 7B: Diffusion Large Language Models*. arXiv preprint arXiv:2508.15487

[65] Bie, T., et al. (2025). *LLaDA2.0: Scaling Up Diffusion Language Models to 100B*. arXiv preprint arXiv:2512.15745

[66] Liquid AI Team. (2024). *Liquid Foundation Models (LFM)*. https://www.liquid.ai/blog/liquid-foundation-models-v2-our-second-series-of-generative-ai-models

[67] Dell'Acqua, F., et al. (2023). *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*. Harvard Business School Working Paper 24-013.

[68] Naskręcki, B., et al. *Mathematical discovery in the age of artificial intelligence*. Nature Physics (2025). https://www.nature.com/articles/s41567-025-03042-0

[69] Meta *Meta Announces Nuclear Energy Projects, Unlocking Up to 6.6 GW to Power American Leadership in AI Innovation*. Meta (2026). https://about.fb.com/news/2026/01/meta-nuclear-energy-projects-power-american-ai-leadership/

[70] Wikipedia. *Colossus (supercomputer)*. https://en.wikipedia.org/wiki/Colossus_(supercomputer)

[71] Crosley, B., (2026)*xAI Colossus Hits 2 GW: 555,000 GPUs, $18B, Largest AI Site* https://introl.com/blog/xai-colossus-2-gigawatt-expansion-555k-gpus-january-2026