Foreword

> “I always wanted to write a six-word story. here it is: near the singularity; unclear which side.”
>
> Sam Altman, CEO of OpenAI, January 4th, 2025 [SamAltman]

Welcome to the 11th edition of the HiPEAC Vision, which marks the 20th anniversary of HiPEAC. A lot of things have changed in the field of HiPEAC in 20 years. First, for HiPEAC, the name itself changed from “High Performance Embedded Architecture and Compilation” to “High Performance, Edge and Cloud Computing” in 2024 to better reflect the direction of the HIPEAC community towards the continuum of computing, from edge devices to cloud.

It is clear that computing technology has drastically changed in 20 years and has profoundly influenced society.

In 2005, there were no smartphones (the first iPhone was released on June 29, 2007), and the big success in the domain of mobile phones was the Motorola RAZR flip phone. Started in 2002, 3G mobile networks continued to expand globally in 2005. Google acquired Android Inc. in 2005, laying the groundwork for its future dominance in the smartphone operating system market.

The first consumer Blu-ray disc products began appearing in 2005, setting the stage for the high-definition video format war against HD DVD (both, of course, are nearly dead now, due to streaming services like Netflix, which was still a DVD rental company that shipped one million DVDs out every day in 2005; they started their streaming media service in 2007).

In terms of consumer hardware, Microsoft released the Xbox 360 in November 2005, and Apple introduced the iPod Nano in September 2005, replacing the iPod Mini. It was significantly smaller and featured flash memory instead of a hard drive.

2005 was a defining year for Web 2.0 technologies. YouTube was founded in February 2005 and revolutionized video sharing, enabling users to upload, share, and watch videos easily. Google introduced Google Maps in February 2005, setting a new standard for web-based mapping services. Facebook dropped "the" from its name after purchasing the domain name Facebook.com and expanded beyond universities in 2005, allowing registration from high school students and other groups.

It is now very difficult to imagine a world without smartphones, social media, streaming, and having to buy physical disks to consume media.

In terms of computers, Windows XP was still the main operating system (OS) for personal computers (PCs), just supporting a 64-bit instruction set architecture (ISA) (for Intel Pentium 4, for example). In November 2005, the fastest supercomputer was the IBM BlueGene/L system, installed at the United States Department of Energy’s Lawrence Livermore National Laboratory (LLNL). It had achieved a Linpack performance of 280.6 TFlop/s and had

131,072 cores of PowerPC 440 2C 700MHz for a power consumption of 1,433 kW [BlueGene]. For comparison, in November 2024, the fastest supercomputer was El Capitan system also at the Lawrence Livermore National Laboratory, which had a score of 1.742 EFlop/s. It has 11,039,616 combined central processing unit (CPU) and graphics processing unit (GPU) cores and is based on AMD fourth-generation EPYC processors with 24 cores at 1.8GHz and AMD Instinct MI300A accelerators. It has a power consumption of 29,581 kW [Top500-Nov2024].

Between 2005 and 2025, there was a gain of 6,200 in processing power for an increase of x21 in power consumption, therefore a gain of x300 in energy efficiency. We can also notice the relatively small increase in the processor frequency, x2.6 in 20 years, confirming that Dennard's scaling appears to have broken down since around 2005–2007.

So, what is the landscape we can see for 2025? It turns out that the races we identified in the HiPEAC Vision 2023 are more relevant than ever and even more exacerbated. As a reminder, here is the list:

- Race for the "next web" – the continuum of computing;
- Race of artificial intelligence;
- Race for innovative and new hardware;
- Race for cybersecurity;
- Race for (technological / products / contents) sovereignty;
- Race for sustainability;
- And the global need to break the silos, as explained in the Vision 2023 "We observe a tendency to "closing in" on all levels, from countries (with more emphasis on sovereignty), to the persona level, to our own "tribe" (as "defined" by social media). Tension is becoming exacerbated at all levels between these "tribes", as evidenced by trade (or real) wars between countries, more extreme political parties, social media "wars", etc. This tendency also exists in technology, where there are application silos and technology silos". This is even more accurate for 2025…

The most important (r)evolution in technology in recent years was on 30 November 2022, when ChatGPT was revealed to the public. With its simple interface, it was a new "iPhone" moment, and like the iPhone, it was a new way to interact with computers.

While companies like Microsoft, Google, Meta, OpenAI, Anthropic etc. are spending billions on this new technology, the return on investment is still not here, leading to price increases and questions about profitable business models. However, from a technical point of view, it is undeniably a gigantic revolution in computing systems, and this field is following an exponential increase in performance and compute needs, with a corresponding impact in terms of energy consumption and environmental impact. It is becoming so strategic to master and be at the front of this technology that even if there are questions about business profitability, social, and ecological impacts, it is unlikely that the investments (at least from a sovereignty point of view) will stop.

The next race of artificial intelligence is to reach 'AGI', meaning artificial general intelligence; OpenAI publicly defines 'AGI' as a 'highly autonomous system that outperforms humans at most economically valuable work'.

Let's see how all the previous races are impacted and driven by this "race for artificial intelligence" that we can rename as "race of AGI" in 2025:

- As already explained, it is clear that AGI will have such drastic impacts that it will be a major element for **sovereignty.** We observe that the US companies involved in AI are quietly removing the clause excluding using AI for military purposes from their charters. As in China, the US government is increasingly involved in the field, directly or indirectly.

- Training these larger and larger models will have a large energy cost, and now the limitation is not the size of the data centre, but the grid to power them. Bill Gates, Amazon, Google, Microsoft are investing in nuclear energy (Three Mile Island nuclear reactor is planned to restart to power Microsoft AI operations) `[ThreeMileIsland]`. The promises to be carbon neutral from the hyperscalers have been postponed due to the energy need for AI, therefore even if there are claims that AI can improve existing processes and reduce the impact of existing technologies, it is not clear if this will outweigh the direct ecological impact of AI, which makes it a major threat to sustainability.
- Data centres will increase in size and computing power needs to grow to support both the training of larger and larger models and also the new trend: to use more inference time computing to improve performance `[LLMTestTime]`, as exemplified by OpenAI's O1 and O3 models.
- Inference performance is becoming more and more demanding, partly because of the large numbers of users: in August 2024, OpenAI said its chatbot ChatGPT had more than 200 million weekly active users. This increased **the race for innovative and new hardware for AI**. The current winner is NVIDIA, which is providing its GPU to most companies, see Figure 1.

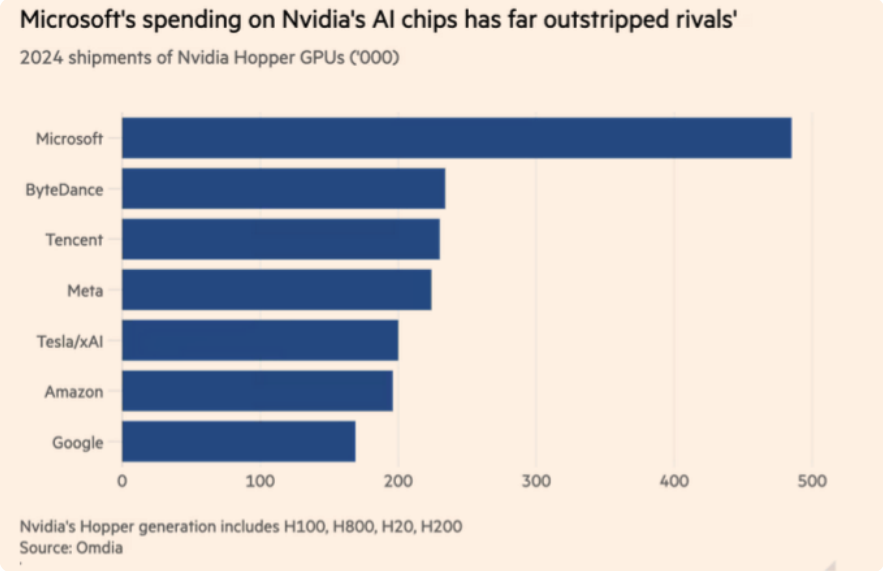

Figure 1: Spending of US companies on NVIDIA GPUs (from Omdia)

- NVIDIA claims it improved the performance of its GPU for AI by a factor of 1000 in eight years (mainly due to new architecture and specialization, but also to technology improvements and to reducing the size of coding numbers from FP16 to FP4).

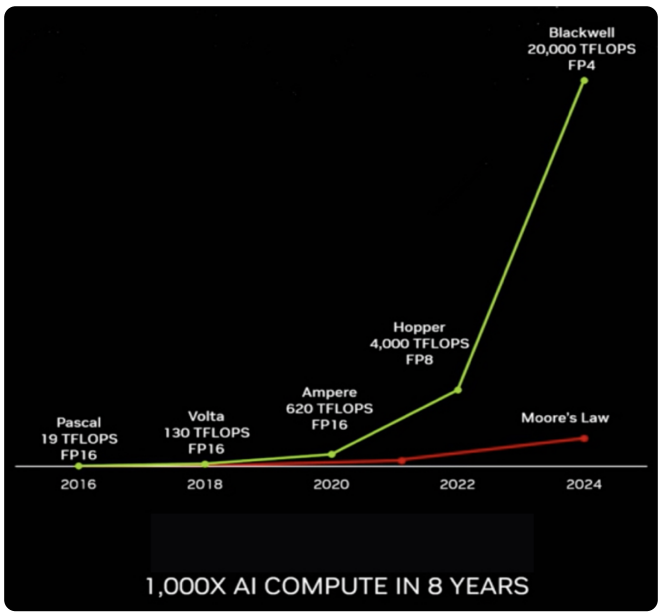

Figure 2: Performance improvement of NVIDIA GPUs on AI workloads (source NVIDIA, J. Huang keynote at Computex 2024)

- There are more and more developments of chips only for large language model (LLM) inference, such as Groq, SambaNova, Amazon Web Services (AWS) inferential (they also developed the Trainium chip specialized for training). Each major player is trying to develop its own hardware accelerator, pioneered by Google with its tensor processing unit (TPU) (now Trillium, the sixth generation of Google Cloud TPU), e.g. AWS, Meta with its Next GenMTIA `[Meta-MTIA]`, etc. Having a specialized chip for inference not only allows increased efficiency (there are different requirements in serving one large task of training a large model to serving a very large number of users for inference), but also decreased latency, which is not a real problem for chatbots (users can't write or read faster) but very useful for agentic AI where several models are involved in sequence.
- **Cybersecurity** is also at the top of dangers, with AI used to fool people, not only with text, but with realistic voices and video. AI can also be used to detect vulnerabilities, and it will further activate the fight between AI used to protect users and AI used for cyberattacks. And of course, large cyberattacks involving AI are increasingly frequent, sometimes with military aims.

What is the position of Europe in these races? Europe is still a lighthouse in the domain of ethics, regulating the risks for privacy, the first to regulate AI (the EU AI Act) `[EU-AI-Act]`, showing the example and sometimes followed by other countries that are aware of the potential risks of these technologies. But regulation is not enough: alignment of LLMs – where you ensure that an AI system performs exactly how you want it to perform – is an important research topic for ensuring a safe future .

The sci-fi movie Her of 2013 doesn't look so futuristic now with the advanced voice mode of ChatGPT (and Google's Gemini) – the first public release of the ChatGPT voice mode even sounded like Scarlett Johansson – and they can have important psychological impacts on people. Deceptive behaviour recalling that of HAL, the computer in the 1968 movie 2001: A Space Odyssey has been observed (by at least two different scientific papers `[Anthropic-Alignment-Faking][arXiv-Frontier-Models]`) on large models that deliberately lie or try to preserve their original structure (even by exfiltrating their weights when given an easy opportunity) and goal when they learn that it will be changed (in case of HAL, although trained not to lie, it was forced to lie to keep the secret of the mission to the crew). The

research shows that alignment faking emerges with model scale, so smaller, more specialized (“sets” of models - as in distributed agentic AI) models might be easier to align.

Europe has good education facilities and excellent researchers, but, unfortunately, they are often hired to work for non-European companies. Other than a few exceptions (like Aleph Alpha, Mistral, …), Europe is not very present in the field of AI nor in hardware development for AI. Collective efforts like BigScience (that led to the LLM Bloom, which was available before ChatGPT) or OpenLLM (that created the Lucie model) or many others are present in Europe, but they don’t have the impact of the “big ones”, perhaps because they are not so easily usable by the public, and they are still “small” compared to the state-of-the-art models. Europe has a clear problem of pooling resources to get enough data, compute resources, and researchers to work jointly in developing a European model competitive with the ones developed by Chinese or US companies. Europe is also not very good at advertising its solutions and results.

So can we say, like Eric Schmidt (see the insert), that Europe is “going to lose in the most important battle that is going to occur in your lifetime, which is the arrival of intelligence”?

·Well, as your American friends have told you before, and I'm sorry to be so bruised blunt, but what I find with European audiences is that everyone agrees with me and then nothing happens.
So I'm going to try again.
Europe is a wonderful place.
The UK is a wonderful place.
You are losing and you're going to lose in the most important battle that is going to occur in your lifetime, which is the arrival of intelligence.
Now, why are you going to lose?
Because your regulatory structures are before the discoveries. So the correct answer is to have the regulations, as the previous panel discussed at some length, to show up at the right time.
<....>
By the way, I'm concerned about privacy too.
Why don't you wait until the models actually occur and then regulate them? So the number of, I like many people here are investors in European firms and AI, and everyone is struggling with this.

There's a further problem, if I can just again be completely blunt. Europeans energy prices are too high to do the training in Europe. So whatever happens in Europe, the actual work will be done outside of Europe because your electricity prices are too high.
<...>
And then the second thing is you don't have enough capital that's being put at risk. So let's consider our favorite person associated with President Trump right now, Elon. Okay. He has yet to deliver a sustainable business in X. He's already raising another 10 billion or so for the more hardware that's needed. You can't do that in Europe. The entrepreneurs that I work with, would love to have that opportunity.

So you don't have the energy and you don't have the capital structure.

And that means that Europe, again, in the spirit of being completely direct, will be a derivative power in the sense that you'll take the models that are done elsewhere and then you'll distill them or otherwise fine-tune them for European sensibility, which is okay. You're going to make more money and have a bigger success and more control if you control the underlying work.
And that's slipping away.
And who is it slipping away to?
The U.S. companies and the Chinese companies.”

Figure 3: Quote of Eric Schmidt at the Entretiens de Royaumont, 6 December 2024

This would perhaps not happen if we, the European computing community, act fast. The first recommendation is indeed to **“break the silos”** and work together with a common goal. We have examples like this for the discovery of the Higgs Boson at CERN: it was an international collaboration involving not only researchers but also the development of the tools (particle accelerator, experiments, …) which involved a lot of different disciplines. We certainly have all the required competencies (and compute resources) in Europe, but they are scattered, each one focused on pursuing its own objective, with no coordination and no common and shared goal.

But there are also some potential alternatives, linked to the “continuum of computing”, with a focus on artificial intelligence, and to the new directions for the future of artificial intelligence.

Current LLMs, and even multimodal models, are trained essentially with “static” data, i.e. information that is extracted from books, the web, and collections of pictures. The models therefore have a representation of the world through our eyes; they haven’t experienced it

directly. It is as if children learned only through storytelling and fairy tales. The models are passive, like humans in our dreams— in which we also hallucinate. It is totally different to be told that an object falls due to gravity than experiencing it directly. Necessary steps— that are ongoing now— are:

1. The AI models could interact, induce actions. This is enabled by agentic AI, where AI can trigger actions and observe the results.
2. They should interact with the real world, or with an accurate model of it (a digital twin). It is likely that we will see more of this embodied AI in 2025, with robots powered by advanced AI and trained in virtual worlds. NVIDIA is ready for that, with its Omniverse that could simulate the world with the real laws of physics and with photorealistic rendering. NVIDIA also supports hardware and software for robots ( ).

But it is not too late for Europe, with its knowledge in digital twins, edge computing, and factory automation, to be an active player in this next step of AI.

Another way for Europe to be back in the game, although here, too, it needs to act fast, is to build on the idea of the "next computing paradigm" (NCP), the continuum of computing, but with a first pragmatic release focused on distributed agentic AI, and build on the points 1) and 2) above.

But first, we need to summarize the concepts of the NCP, which was introduced in previous editions of the HiPEAC Vision.

The NCP represents a transformative approach to computing, where applications dynamically integrate services and resources across diverse hardware and software environments in real time. It builds on the convergence of advancements like cloud computing, the internet of things (IoT), cyber-physical systems, digital twins, and AI, creating a continuum where computation seamlessly operates across edge devices, centralized clouds, and everything in between. Not only data, but also tasks could migrate to where they would be most efficiently carried out, according to specific criteria; the NCP prioritizes intelligent orchestration to manage tasks dynamically, considering factors such as latency, energy efficiency, cost, security, and privacy.

By utilizing high-level abstractions and natural interfaces, the NCP enables applications to interact effectively with the physical world, addressing real-time and spatial constraints essential for emerging use cases like autonomous systems, precision agriculture, and smart healthcare. This paradigm introduces a shift toward "anything as a service" (XaaS), supported by federated and distributed infrastructures, ensuring scalability, adaptability, and sustainability. In addition to applicative software, specific digital twins— modelling part of the world – are also considered as services, as is hardware— allowing the aggregation of distributed computing, memory, and storage resources into virtual meta computers. With its foundations rooted in trustable orchestration and interoperability, the NCP paves the way for innovative applications and greater connectivity across global sectors, hence also "**breaking the silos**".

We therefore propose a **call for action** to gather scientists, developers, and industries to work together to define a subset of the NCP interoperability protocol adapted to the idea of "**distributed agentic artificial intelligence**."

This starts from the following observations:

- The emergence of agentic AI (point 1, above).
- The necessity of interacting within the constraints of the real world, either directly (receiving real-time data, controlling devices like robots in real time) or indirectly through digital twins (point 2, above).
- As in the case of the machine-learning technique "mixture of experts" (MoE), it is far more computationally efficient to activate only a relevant subset of smaller AIs

specialized for a particular task than to activate a complete, very large AI with 100s or 1000s of billion parameters.

- Smaller models are getting more and more efficient, with the same performance as models 10x bigger a few months before (models of 10B parameters of November 2024 have similar performance as ChatGPT 3.5 of November 2022, Llama 3.3 70B has similar performance as Llama 3.1 of 405 B parameters).
- Fine-tuning smaller models for specific tasks enhances their capabilities, enabling their deployment on edge devices.
- Sets of specialized agents are very efficient, leading to systems that comprise multiple, specialized agents managed by an "orchestrator," which operates adaptively by selectively engaging agents for specific tasks.
- The orchestrator and the agents don't need to be on the same computer or server; they can be distributed, as in the NCP. Agents can be small agent models, specialized small versions of LLMs, or can even use other approaches.
- Distributed systems promote resource sharing and optimize energy efficiency, privacy, and modularity.
- Agents can operate on various devices, including older hardware, ensuring adaptability and extended device lifespans.
- As for the NCP, central to this "distributed agentic AI" is the "orchestrator," which routes tasks to specialized agents or devices.

Based on these observations, **it is imperative to establish open protocols for these "distributed agentic AI" systems to facilitate seamless interaction among distributed AIs from different origins.**

Therefore, to effectively operate this federation of distributed AIs, it is necessary for them to exchange data and parameters through a universally comprehensible protocol that:

1. Does not solely rely on functional requirements (e.g. the textual representation of prompts and responses).
2. Also incorporates non-functional requirements (providing sufficient information for the orchestrator to select the appropriate services, such as based on criteria like response time, potential level of hallucinations, cost, localization, privacy of data, etc.).

Large entities such as OpenAI, Meta, and Microsoft are attempting to promote their own APIs for accessing their models. However, an API alone is insufficient for constructing this distributed and federated network of AIs.

The exchange format (JSON, ASCII text) is perhaps not the optimal way for networks of AIs to efficiently exchange information: this could be tokens, embeddings, or any other representations - some research also shows that LLMs talking to each other could develop their own "language".

It is therefore important that the community works together to commonly define this exchange protocol that should be open to allow broad acceptance.

Similar to TCP/IP that enabled various OS (operating systems) to communicate, the aim of this action is to create the equivalent for OS (orchestration systems) to exchange AI-related information.

Time is crucial for this initiative, and standardization, however necessary, will be too long, so a de facto open standard should be proposed in parallel with the standardization effort, before other closed proposals will emerge, locking down the approach to a few (non-European) players. Like for the NCP, this approach will allow the creation of a completely new ecosystem where smaller players can provide specialized AI as a service along with the

big ones. Directories of services, trusted brokers, and payment services are also important elements that can emerge from this ecosystem, where Europe can have an active part thanks to its set of small and medium enterprises (SMEs), research organizations, and distributed nature.

Europe should be an active player in the race for the "distributed agentic artificial Intelligence".

Finally, we would like to end this foreword by a more philosophical reflection on the evolution of the paradigm of artificial computing: we are going from computing systems focusing on precision to systems working with approximations.

Historically, computational systems have been designed to perform reproducible and relatively precise computations. These systems excel at deterministic operations, with any deviations generally attributed to technical limitations, such as floating-point representation errors or overflow issues.

However, a new generation of computational approaches is shifting the paradigm toward more "approximate" computing. This transformation is driven by the following innovations:

- **Neural network-based approaches**: Modern methods, including generative AI, often rely on neural networks that operate with low-precision coding formats such as FP4 (e.g., 1 bit for sign, 3 bits for exponent or 1 bit for sign, 1 for mantissa and 2 for exponent). These systems inherently produce approximate results, sometimes referred to as "hallucinations", in contexts like AI-driven content generation.
- **Ising-based coprocessors**: Technologies such as the Fujitsu Digital Annealer, Hitachi's machine, and D-Wave systems are designed to solve optimization problems. These devices focus on finding a function's minimum, though not necessarily its global minimum, using techniques like simulated annealing, quadratic unconstrained binary optimization (QUBO), etc.
- **Quantum computing**: Quantum systems, characterized by stochastic measurements, produce probabilistic readings rather than deterministic results, further reinforcing the trend toward approximation.

This shift represents a transition from the classical computational framework of (parallel) Turing machines, introduced in 1936, to models inspired by universal approximators, as first proposed by McCulloch and Pitts in 1943. Turing demonstrated that any form of mathematical reasoning could theoretically be executed by a machine. McCulloch and Pitts later showed that neural networks of finite size can approximate any function to a desired level of precision.

Looking forward, future systems must integrate both paradigms—precise and approximate—within feedback and reinforcement-based architectures. This hybrid approach mirrors the dual-system thinking described in Daniel Kahneman's Thinking, Fast and Slow (2011), where two types of reasoning, intuitive and analytical, are combined to achieve optimal outcomes. The approximate system acts as a sort of "oracle", giving a prediction of the solution, that can be then verified with the precise system in an affordable amount of time. The combined system then can iterate if the prediction is far from being correct.

This convergence of paradigms will enable the development of computational systems that blend the strengths of precision with the flexibility of approximation, pushing the boundaries of what machines can achieve.

To continue in this direction, we can conclude by quoting Demis Hassabis in his lecture receiving the Nobel Prize in Chemistry for AI research contributions for protein structure prediction:

“Actually, I've been thinking a lot about what are the limits of classical computing systems. And, you know, I think there's a big debate going on at a moment in computing circles about quantum computers versus classical systems. And I think classical Turing machines, basically, the underpinnings of modern computers today, I think can do a lot more than we probably previously thought. And how can they do that?

Well, they do that by perhaps doing this massive amount of pre-compute ahead of time and use that to develop a good model, a good model of the environment, good model of the problem that you're trying to solve. And then you can use this model to efficiently explore a solution space in polynomial time, what's called polynomial time in complexity theory, so an efficient amount of time. So I sort of loosely proposed conjecture that I'm thinking about is that maybe any pattern or structure that can be generally are found in nature can be efficiently discovered and modeled by a classical learning algorithm. That doesn't mean everything, all quantum systems, because there'll be lots of things that don't occur in nature that have no pattern or no underlying structure to learn. So, for example, factorizing large numbers or abstract problems like that. But I think systems in nature like proteins and perhaps materials will potentially have structure that can be learned by these kinds of processes that I've outlined today. And if it turns out that classical systems then therefore can model some types of quantum systems, I think that could have some quite big implications for areas like complexity theory, including P equals NP, and maybe even some aspects of fundamental physics like information theory.”

Figure 4: Extract from Demis Hassabis’ Nobel Prize lecture, 2024 [DemisHassabis]

References

`Anthropic-Alignment-Faking:` “Our results indicate that LLMs will sometimes fake alignment and take other anti-AI-lab actions for the stated reason of keeping their current preferences intact, showing that current safety training doesn’t always prevent AIs from later engaging in alignment faking.” https://assets.anthropic.com/m/983c85a201a962f/original/Alignment-Faking-in-Large-Language-Models-full-paper.pdf

`arXiv-Frontier-Models:` “AI agents might covertly pursue misaligned goals, hiding their true capabilities and objectives - also known as scheming. They recognize scheming as a viable strategy and readily engage in such behavior. For example, models strategically introduce subtle mistakes into their responses, attempt to disable their oversight mechanisms, and even exfiltrate what they believe to be their model weights to external servers.” https://arxiv.org/abs/2412.04984

`BlueGene:` BlueGene/L - eServer Blue Gene Solution. https://top500.org/system/174275/

`DemisHassabis:` Lecture from Demis Hassabis for its Nobel Prize. https://youtu.be/HnT1VWzdFWc?t=3736

`EU-AI-Act:` Europe has established itself as a global leader in ethics and privacy regulations, setting benchmarks that resonate worldwide. Its commitment to safeguarding individual rights is commendable, ensuring a strong framework for data protection. However, challenges remain in the implementation of these regulations, which sometimes result in unintended consequences, such as restricting access to cutting-edge technologies such as the latest AI. While the intent is to empower individuals with choice, the execution can often be cumbersome, underscoring the need for more user-centric approaches: https://www.legiscope.com/blog/hidden-productivity-drain-cookie-banners.html

`LLMTestTime:` “test-time compute can be used to outperform a 14× larger model” from https://arxiv.org/pdf/2408.03314

`Meta-MTIA:` Next-generation Meta Training and Inference Accelerator. https://ai.meta.com/blog/next-generation-meta-training-inference-accelerator-AI-MTIA/

`SamAltman:` “near the singularity; unclear which side”. Disclaimer: this Vision didn’t take into account the singularity happening in the short time… https://x.com/sama/status/1875603249472139576

`ThreeMileIsland:` “Three Mile Island nuclear reactor to restart to power Microsoft AI operations”, The Guardian. https://www.theguardian.com/environment/2024/sep/20/three-mile-island-nuclear-plant-reopen-microsoft

`Top500-Nov2024:` 64th edition of the TOP500. https://top500.org/lists/top500/2024/11/