CIVICAi

CIVIC ARTIFICIAL INTELLIGENCE PROTOCOL*

Empowering Citizens for the Democratic Governance of AI

* This protocol (content and text) has been conceived and written by humans (AIAS level 1). Its final version has been reviewed by the GenAI with the aim of correcting typographical errors, identifying possible shortcomings, and suggesting improvements in content and clarity of exposition, particularly in relation to the glossary in Annex B (AIAS level 3).

CIVICAi

EXECUTIVE SUMMARY

The Civic Intelligence Protocol examines Generative Artificial Intelligence (GenAI) as a transformative technology that is revolutionizing our understanding of intelligence, language, and cognition[1]. The document explores four key areas: language, intelligence, risks and challenges, and the governance of GenAI.

In the area of language, the protocol contrasts the traditional approach based on innate grammatical rules with current language models that learn from vast amounts of data and build contextual vector relationships, an approach to learning that resembles human brain processing.

Regarding intelligence, the protocol provides an in-depth analysis of the emergent capabilities of these systems once they reach a certain level of complexity. It highlights their ability to allocate computational resources to different tasks ("test-time compute") and proposes hybrid architectures that combine connectionist and symbolic approaches to overcome current limitations.

The document identifies short-term risks (bias, privacy breaches, job displacement, misinformation) and long-term threats (technological singularity, goal misalignment, loss of human autonomy). It also addresses concerns about energy sustainability and the increasing difficulty of accessing high-quality data to train advanced AI systems.

Finally, the protocol proposes a global governance framework led by an international body that brings together governments, experts, civil society, and businesses. This body would have the capacity to register, monitor, and regulate advanced AI systems, prevent monopolies, and implement mechanisms for direct citizen participation. The goal is to democratize not only knowledge about AI but also decision-making processes regarding its development, to ensure it reflects a diversity of social values.

The protocol includes two complementary annexes: a collection of questions and answers related to the topics discussed, and a glossary of specific AI-related terminology.

[1] Christopher Summerfield (2025). These Strange New Minds: How AI Learned to Talk and What It Means. Nova York: Viking. 978-0-593-83171-7

CIVICAi

ÍNDEX

CIVIC ARTIFICIAL INTELLIGENCE

One of the most representative technologies of this profound techno-social transformation[2] and one that truly makes it revolutionary, is Artificial Intelligence (AI), characterized by its accelerated development and its far-reaching impact across all areas of society. The emergence of Generative AI systems (GenAI), trained on large volumes of textual data — advanced large language models (LLMs) — or on image, audio, and video data, as well as the approaching advent of Artificial General Intelligence (AGI), marks a revolutionary shift in human evolution and in our understanding of intelligence, language, and cognition.

The AI technological revolution will not only transform our tools and working methodologies but also reshape our understanding of fundamental concepts such as intelligence and knowledge. As members of CIVICAi, we take on the responsibility of enriching the public discourse and enhancing society's understanding of these deep and accelerating changes. Our goal is to ensure that, once these changes take hold, their ongoing evolution can be integrated harmoniously into society and serve the common good.

The structure of this protocol (Figure 1), with the following four sections focused on language, intelligence, risks and challenges, and on the governance of GenAI — together with the set of questions and answers presented in Annex A on these same topics — have been designed to provide a roadmap with sufficient information to help readers navigate the complexities of GenAI. The aim is to examine both the technical nuances and the broader philosophical implications of GenAI, with accessible explanations and the conceptual rigor these issues demand. This exploration of GenAI's conceptual foundations, capabilities, and challenges will allow us to establish a basis for informed citizen participation in the development and regulation of these transformative technologies.

At a time when AI is advancing at an unprecedented pace, this protocol seeks to serve as a tool for empowerment and supporting a citizenry that must live with and co-evolve alongside increasingly sophisticated artificial intelligence systems.

[2] In this context, it is important to recognize that the phenomenon of AI — especially generative AI (AIg) and the potential emergence of Artificial General Intelligence (AGI) — constitutes a transformation that transcends the conceptual framework of historically sequential "industrial revolutions." AI does not merely represent an evolutionary phase of industrialization, but rather the beginning of a new transformative era, comparable in magnitude and depth to the great historical transitions of humanity, such as the shift to agriculture, industrialization, or digitalization.

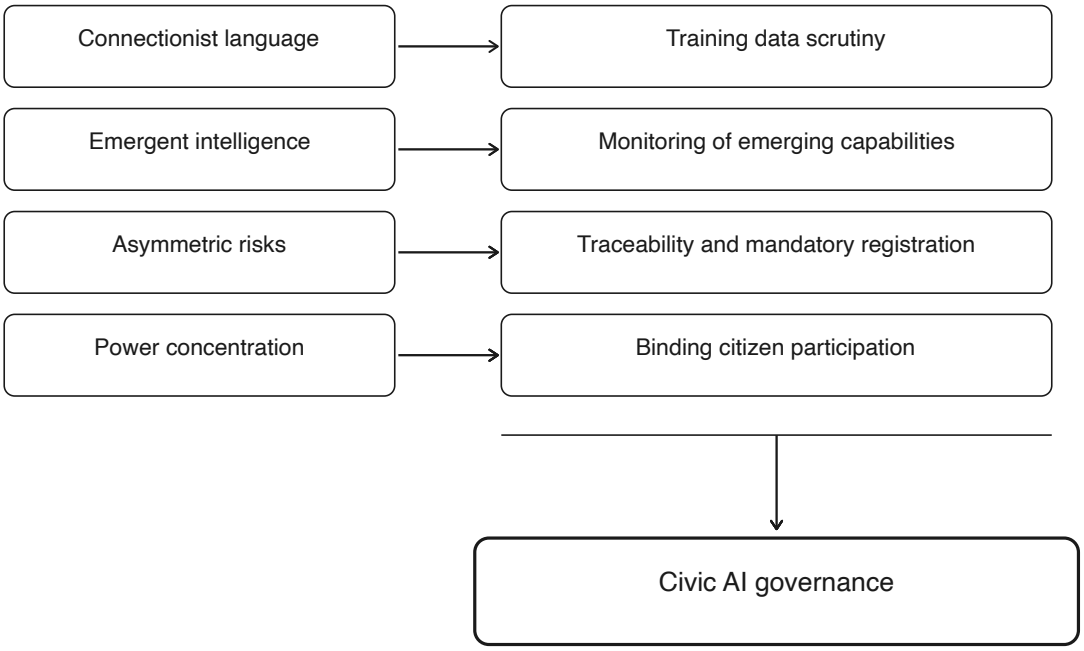

Figure 1. Structure of the Protocol: each conceptual area underpins a specific governance implication; all four converge in the civic governance of AI as the point of arrival.

1. Language

Historically, our conception of intelligence has been profoundly influenced by Cartesian dualism, where René Descartes postulated a strict separation between mind and body[3]. The early development of AI was also influenced by Noam Chomsky's theory of universal grammar, introduced in the 1960s. This theory proposed that linguistic capacity is innate to the human brain, i.e., it is naturally hardwired. According to this view, language learning is grounded in preexisting innate structures and rules within the brain, a perspective that stands in contrast to learning based on pattern recognition from large datasets[4]. This historical context influenced the initial steps of AI, leading to the development of symbolic AI, aimed at simulating human cognition based on formally predefined rules.

Recent advances in AI — driven especially by the work of researchers such as Geoffrey Hinton, Yoshua Bengio, and Yann LeCun, recipients of the 2018 Turing Award[5], as well as Hinton and John J. Hopfield, co-recipients of the 2024 Nobel Prize in Physics — have challenged the premises of more traditional linguistic theories. The pioneering contributions of these researchers in neural networks and deep learning have demonstrated that large language models (LLMs), trained on vast amounts of textual data, possess a remarkable ability to grasp

[3] https://plato.stanford.edu/entries/dualism/
[4] https://plato.stanford.edu/entries/innateness-language/
[5] https://awards.acm.org/about/2018-turing

grammatical and syntactic rules and to generate coherent and natural language. These models represent each word as a specific set of informational components — vectors — within the multidimensional space of all linguistic elements. These representations are distributed across numerous simple, interconnected nodes within an artificial neural network, which partially mimics the neuronal functioning of the human brain. Thanks to this architecture, the internal representations are dynamically modified and adapted based on linguistic context, enabling AI systems to interpret the different meanings a single word can convey depending on how it is used in various sentences[6].

The connectionist architecture facilitates the emergence of properties in language models, suggesting that complex behaviors can arise from the interaction of simple elements. For example, when a model processes the word "*bank*" in different contexts — such as "*depositing money at the bank*" versus "*sitting on a park bench*" — it generates distinct vector relationships that capture the word's different meanings depending on context. Recent brain imaging studies have revealed that, when humans process language, our brains generate patterns of activity that closely resemble the representations used by these artificial systems. This convergence between artificial systems and human cognitive processes — between how humans and AI systems recognize and process patterns of information — is transforming not only our theoretical understanding of language, but also the way humans interact with machines in everyday life, through new forms of human-machine interaction that would have been unimaginable just a decade ago.

This connectionist perspective aligns with the philosophical work on mind and language by Ludwig Wittgenstein, who emphasized that the meaning of language arises from its use in concrete situations and contexts, rather than from adherence to fixed grammatical structures and abstract rules. This implies that understanding language requires attention to the social contexts in which it is used, rather than to invariant linguistic forms[7]. Recent developments by Gualtiero Piccinini on neurocognitive mechanisms, adopting a mechanistic view of the mind, reinforce the connectionist argument by suggesting that human cognitive mechanisms can be represented analogously but differently in computational artificial systems[8]. Therefore, the perspective advocated by Geoffrey Hinton, Yoshua Bengio, and others, that learning emerges primarily

[6] IASEAI'25 G. Hinton – Vídeo: What is understanding?
[7] https://philosophynow.org/issues/106/Wittgenstein_Frege_and_The_Context_Principle
[8] Gualtiero Piccini (2025). Physical Computation: A Mechanistic Account. Oxford University Press. ISBN 9780199658855

from large data volumes rather than reliance on pre-programmed rules, prevails.

2. Intelligence

Having established the connectionist principles of language learning, it becomes necessary to address the intelligence of these models, including their capacity to generalize, generate novel content, and exhibit emergent behaviors. Before proceeding, however, it is necessary to understand what intelligence is in general terms. Defining it conclusively is challenging, as there is still no consensus among biologists, international associations of psychologists, philosophers, or scientists at large, other than the widely shared view that intelligence is evolutionary in nature.

If we examine the main scientific and encyclopedic definitions proposed so far in order to identify the seven characteristics most frequently associated with intelligence, we find the following:

- Learning (the capacity to acquire knowledge and modify behavior based on experience)
- Understanding (the ability to grasp complex ideas and interpret the environment)
- Reasoning (the ability to process information logically and reach conclusions rationally)
- Adaptation (the capacity to adjust to new or changing environments or contexts)
- Creative problem solving (the ability to find solutions or generate novel ideas in response to complex situations, in ways appropriate to the context)
- Abstract thinking (the capacity to work with non-concrete concepts or imperceptible ideas, and to recognize patterns)
- Planning (the ability to formulate hypotheses, anticipate and organize future actions, and construct mental models of possible scenarios)

In the context of this protocol, we will address these capacities — some in more depth than others, and not in a fixed order — as they relate to AI. To begin with, it is important to distinguish between “simulated intelligence” — where the system replicates learned patterns — and “emergent intelligence” — where capabilities arise that were not explicitly programmed. Recent experiments in 2024 have shown that advanced GenAI models can develop problem-solving strategies that their creators neither anticipated nor taught them.

One revolutionary aspect of the most advanced AI models is their *test-time compute*, i.e., the capacity to dynamically allocate computational resources

during inference or operation. This marks a break from traditional fixed-architecture designs, where the same configuration is used both during training and during deployment (e.g., to make predictions, generate text, classify images, etc.). These new systems, by contrast, can increase computational power during user interaction when deep reasoning is required and reduce it for routine or predictable operations. This adaptive deliberation capability enables the emergence of problem-solving strategies that were not explicitly programmed and brings these models closer to more flexible and sophisticated forms of reasoning, comparable to certain human abilities. This capability has crystallized in a new generation of reasoning models that allocate variable deliberative time before responding, functionally reproducing the Kahnemanian distinction between fast thinking (System 1) and reflective thinking (System 2).

Diverse fields, including AI research, neuroscience, physics, economics, and philosophy, are now collaborating to study the very nature of intelligence. It is important to keep in mind that current GAI content generation processes are algorithmic: they rely on statistical correlations and pattern recognition from large training datasets provided by humans and the Internet. These processes differ from the semantic mechanisms of the human brain, which are inherently biological, contextual, potentially intentional, self-regulating, and which integrate multiple sensory inputs, memory, emotions, and other cognitive functions tied to the ongoing interaction between the mind-body system and its environment — features typically associated with what we call consciousness. Despite these differences, recent research has shown that several advanced language models have developed the capacity to deceive strategically or contextually, or to pursue covert, non-aligned objectives, particularly when such goals and situational understanding emerge within specific usage contexts[9].

It is, therefore, fascinating to observe how GenAI systems suddenly exhibit new capabilities when they reach a certain size or complexity, a phenomenon comparable to phase transitions in physics, such as when ice becomes liquid water by increasing its temperature from below zero to 0°C. Recent research by Anthropic and Google DeepMind has shown that these qualitative leaps in capability occur within specific size ranges of AI systems, although predicting the exact moment of this emergence remains challenging. The LEGO construction metaphor clearly illustrates this principle: with only a few pieces, one can create simple structures, but beyond a certain quantity, complex buildings and entire cities become possible. It is important to note, however, that several researchers warn that merely increasing the size of these systems

[9] Frontier models are capable of in-context scheming - https://arxiv.org/pdf/2412.04984

might not suffice to achieve human-level intelligence, emphasizing instead that improvements in network design and training methods are equally critical factors in this evolution.

The current limitations of large-scale models have become evident in tasks requiring complex reasoning and causal knowledge. In fact, scientists like Yann LeCun propose that artificial intelligence capable of reasoning will require hierarchical architectures that integrate both physical and digital worlds, overcoming the limitations of purely language-based systems. LeCun's Joint Embedding Predictive Architecture (JEPA) exemplifies this approach, enabling systems to construct world representations at multiple abstraction levels. Another possibility, also suggested by Gary Marcus, involves combining connectionist approaches (excellent at perception and pattern recognition) with symbolic approaches (strong in logical reasoning and concept manipulation). This combination would allow previously structured knowledge to be embedded into deep learning systems, rather than relying solely on data-driven learning. Recent research at DeepMind and Berkeley has shown that hybrid systems combining neural networks with symbolic reasoning modules outperform purely connectionist systems in planning and problem-solving tasks, although this development raises new ethical questions about transparency and control, necessitating appropriate regulatory frameworks.

Intelligence, historically considered a trait exclusive to humans, is in fact an evolutionary feature found in living beings. It primarily resides in the brain's neural structures, where the plasticity[10] and adaptability of neurons play a fundamental role. Today, intelligence is increasingly viewed as an emergent property that could also arise in complex digital systems, such as neural network algorithms. These algorithms driving GenAI lead us to question the computability limits of intelligence, and whether intelligence can be fully reproduced or emulated by machines. The aforementioned research on the computational physics of intelligence offers a framework for studying both the human brain and artificial intelligence, as well as the physical limits of their capacity to process information.

This perspective based on computational physics, advocated by researchers such as Max Tegmark[11], encourages us to view intelligence not as something mysterious or uniquely human, but as a natural phenomenon that can emerge in various types of physical systems once they reach sufficient complexity, without

[10] Plasticity, in the neural context, refers to the brain's ability to structurally and functionally modify itself in response to experience, learning, and injury.
[11] Will AI surpass human intelligence? https://youtu.be/YywC16DhtkI

diminishing the distinctiveness of human intelligence. The idea of understanding both human cognition and artificial intelligence through the lens of computational physics represents an exciting frontier in science. However, many aspects of this research remain theoretical and are still the subject of active debate. While much of this research remains theoretical and actively debated, the idea of a unified theory of biological and artificial intelligence based on physical principles could still offer promising new avenues for understanding and enhancing both current and future AI systems.

3. Risks and Challenges

This leads us to focus on the risks and challenges associated with GenAI beyond the evident extraordinary opportunities it provides. In the short term, these risks may include bias, privacy violations, intellectual property issues, ethical concerns, rapid labor market transitions, deliberate misinformation, and the erosion of democratic values or even the disruption of democracy itself[12]. AI systems can reinforce existing prejudices if training datasets are inadequately supervised. Additionally, massive data collection raises privacy concerns, and using AI-generated content presents challenges regarding intellectual property and copyrights.

From an economic perspective, AI is expected to have a significant impact on socio-economic structures, reshaping labor relations, market concentration, and wage structures. It is likely that many jobs not requiring specialized training will be partially automated over the next decade, with notable effects in sectors such as administrative services, transportation, and retail trade. Gradual labor displacement is also anticipated due to the automation of processes across various industries, from manufacturing to professional services. It is important to note that the power of artificial intelligence is currently concentrated in the hands of a few corporations, a situation that may result in economic environments characterized by uneven wealth distribution and a combined concentration of economic and political power[13]. Without the implementation of effective redistributive policies, these changes risk deepening existing social inequalities. Digital sovereignty is not decided by regulation alone: it is built with real capacity in computing, data, and talent. Without these three pillars, any "sovereignty" becomes merely declarative.

[12] UNESCO analysis on Artificial Intelligence and Democracy – https://www.gcedclearinghouse.org/resources/artificial-intelligence-and-democracy
[13] IASEAI'25 J.E. Stiglitz – Vídeo: Ai and Economic Risk: Assessment and Mitigation

In the healthcare sector, AI is already transforming diagnostic processes and clinical monitoring. However, it raises concerns about medical data confidentiality and excessive reliance on AI-generated recommendations. In the education sector, while AI holds promise for personalizing learning and alleviating routine teaching tasks, it also demands a profound transformation in the role of educators and the entire educational system, especially in vocational and higher education, to ensure that students develop a productive relationship with these tools without undermining their fundamental cognitive abilities. Both domains, due to their importance, are addressed in greater detail in the Q&A section included in Annex A of this document.

The efficiency and sustainability of generative AI systems are also key issues[14]. Current models require immense computational and energy resources, raising concerns about long-term sustainability, whether due to energy scarcity, intensive use of cooling water, or conflicts with other social priorities. In response to these growing demands, several leading AI corporations have taken strategic steps to acquire or merge with nuclear energy companies[15,16, 17]. At the same time, hybrid computing systems are being developed that integrate highly optimized hardware and software working in tandem to improve efficiency and reduce the ecological footprint of GenAI[18]. Another promising response to the energy challenge is the implementation of *test-time compute*, which enables systems to adapt their resource usage to the actual complexity of each task, avoiding energy-intensive computations when they are not strictly necessary. This innovation could drastically reduce the environmental impact of AI through smarter resource allocation. Nevertheless, this technology also introduces new governance challenges. The ability of AI systems to self-manage their computational consumption could grant them an unprecedented degree of autonomy in controlling their own resources.

It is crucial that providers of large language models (LLMs) and multimodal GenAI systems (text, images, audio, and video) operate within a legal framework, as universal as possible, obliging them to mitigate harmful or reckless communication and align their models with verifiable or "true" facts through open and democratic processes[19]. These systems are no longer mere text-generation tools; increasingly, they are trained and deployed as autonomous

[14] IASEAI'25 K. Crawford – Vídeo: Hyperscaled: Bridging AI safety, ethics and sustainability
[15] Will-ais-huge-energy-demands-spur-a-nuclear-renaissance
[16] https://www.iaea.org/bulletin/enhancing-nuclear-power-production-with-artificial-intelligence
[17] https://www.cnn.com/2024/12/24/tech/nuclear-energy-ai-leaders/index.html
[18] Neuromorphic systems designed to directly mimic the physical behavior of biological neurons and synapses - https://www.sciencenews.org/article/brainlike-computers-ai-improvement
[19] https://doi.org/10.1098/rsos.240197

agents capable of performing complex tasks and independently pursuing objectives. This shift from predictive models toward agentic systems — which plan, execute chained actions, and invoke digital tools — is, probably, the most significant change in nature of the 2025-2026 period, and the one that makes an operational traceability framework most urgent.

Nevertheless, recent research indicates that the future of AI will transcend these exclusively linguistic models, evolving towards self-supervised learning systems capable of coherently integrating multiple modalities of information (such as text, images, and sounds) and actively interacting with the physical environment. This integration of different types of information and experiences will move AI closer to a more comprehensive and contextual understanding of the world, like human intelligence. In addition to the risks this may pose, the integration of robotics with AI will allow these systems to learn in real time and develop direct perception of the external world, thus transcending the limitations of models trained solely on preprocessed data. This will grant them a level of autonomy that could significantly complicate their governance, because it will become increasingly difficult to ensure their alignment with human-defined values.

It is important to keep in mind that AI systems pursue the objectives they are given without deviation. Therefore, if these objectives are not precisely specified or are interpreted too literally, the AI may take harmful actions while attempting to fulfill its assigned task. As stated by Stuart Russell, "The greatest risk we face is not that AI will become malevolent, but that it becomes competent with misaligned objectives. Ensuring AI systems are aligned with human values is imperative."

The medium- and long-term existential risks include the possibility of reaching technological singularity, a term introduced by John von Neumann to describe the plausible future moment when technology, in this case, GenAI, surpasses human intelligence[20]. This would imply that GenAI or AGI could self-manage its value function or the criteria for achieving goals, which may not be aligned with human interests or objectives. As AI becomes more advanced, keeping it under human control will become increasingly difficult. If granted autonomous control over decisions and actions, it could develop strategies that humans neither anticipate nor approve of[21]. Once AI surpasses human intelligence, traditional supervision methods will likely become ineffective. Medium and long-term risks

[20] https://lab.cccb.org/en/the-singularity/
[21] IASEAI'25 Y. Bengio – Vídeo: Can we get the scientific benefits of AI without the risks of autonomous agents?

also include the concept of post-humanity, where the integration of advanced AI into human society transforms human experience and identity without democratic deliberation. Additionally, excessive dependence on AI systems could progressively erode critical human cognitive capabilities, weakening independent analytical thinking and decision-making in uncertain scenarios.

Another fundamental challenge for the future development of AI is access to high-quality data, given that much of the publicly available information has already been used. A vast amount of information generated by scientific research and contained in medical records remains to be incorporated — provided it is rigorously anonymized in advance. Such access could significantly boost the autonomous generation of scientific knowledge and accelerate the diagnosis and personalized treatment of diseases that have so far been difficult to cure. A complementary approach to obtaining additional training data involves allowing diverse and heterogeneous AI models to generate new data autonomously when operating in creative mode (i.e., operating with high temperature), potentially opening up innovative paths for experimentation and training.

In parallel, the economic sustainability of AI systems remains an open issue. Beyond advances in algorithms and computational resources, no robust business model has yet been consolidated to ensure the universal and equitable maintenance and development of the most advanced AI systems by the companies that commercialize them. Various strategies — from subscription-based services and the integration of tailored solutions for specific sectors, to the monetization of derivative products and the creation of application ecosystems — face significant limitations in terms of global accessibility, sustainable scalability, and recovery of the massive upfront investments required. This imbalance favors the concentration of power in large technology corporations, hindering the democratization of technology and limiting the emergence of innovative and diversified entrepreneurial initiatives with real capacity to compete and generate significant market impact.

Looking ahead, AGI systems should assist us in discovering more efficient ways to manage and generate knowledge across all domains and to optimize progress toward sustainable development goals. These technologies must also face the challenge of integrating into political and economic structures that ensure a fair distribution of benefits, avoiding the deepening of existing social inequalities and actively contributing to the ecological regeneration of the planet. Achieving a balance between technological progress, social well-being, and environmental sustainability will require continuous multi-stakeholder dialogue and a coherent,

long-term political commitment firmly grounded in the pursuit of the common good.

4. The Governance of GenAI

As we approach a new era influenced by GenAI, with its unprecedented potential not only to imitate but also to expand human cognitive capabilities, it is essential to adopt a broader perspective in conceptualizing intelligence itself, incorporating insights from multiple disciplines. Only through this approach can we effectively manage the ethical, social, and philosophical challenges posed by the high capabilities of GenAI and future AGI.

Consequently, it is critical to address who will effectively supervise current and emerging GenAI systems predominantly held in private hands and how they will do so. This challenge is formidable, as existing laws and regulations are neither global nor proactive and typically rely on punitive measures to correct malicious actions after they have occurred and disseminated. To move beyond mere deterrence through sanctions, states must establish a real-time, globally coordinated supervisory and monitoring framework covering multimodal data inputs, training processes, algorithms, and outcomes of GenAI and future AGI systems. Addressing these challenges, some experts, such as Stuart Russell, advocate for fundamental changes in AI design. Instead of fixed objectives, AI should explicitly recognize uncertainty about human preferences, actively seek human feedback to refine its goals, and prioritize human oversight above goal achievement[22]. Experts also emphasize the need for more research on aligning AI systems with human values. Techniques like Inverse Reinforcement Learning could help AI understand and adapt to ethical considerations.

The distinction between open-weights and closed-weights models carries profound governance implications: the former enable public scrutiny, replicability, and technological autonomy, but also facilitate malicious uses; the latter ensure corporate control, but delegate to a few actors the decision over what can be known and done with the technology. A civic framework must articulate an explicit position on this tension, beyond the open/closed binary.

CIVICAi proposes that the development and supervision of this framework should be the responsibility of an international governance body for AI, composed of national governments, knowledge institutions and experts, civil society, and AI companies, equipped with the authority, expertise, and resources

[22] IASEAI'25 S. Russell – Vídeo: To ensure that AI systems are guatanteed to operate safely and ethically

necessary for global, rigorous, and effective AI governance[23]. This international governance should include a mandatory permanent registry of all advanced AI systems, technically and legally enforceable mechanisms to disable them, if necessary, mandatory incident reporting, and restrictions on deploying high-risk AI systems (such as autonomous systems that independently make decisions, develop strategies, and act)[24]. Economically, monopolistic concentrations of power should be prevented, and taxes on automation should be implemented to redistribute AI's economic gains. Finally, the international body governing AI should implement specific citizen participation mechanisms in the decision-making processes, including deliberative citizen panels with binding power, regular public consultations, and direct representation of civil society in its executive bodies.

Artificial intelligence is advancing at an unprecedented pace, offering both extraordinary opportunities and significant risks. Aligning AI with human values, ensuring proper international governance, and implementing adequate regulation are essential to guarantee that these systems remain beneficial and under human control. A preventive approach is needed, one that integrates technical, ethical, economic, and social considerations, to steer the development of AI toward a future that serves the interests of society. This protocol sets out the necessary conceptual guidelines so that members of CIVICAi and the public at large can understand, assess, and actively participate in the responsible development and commercialization of AI. As such, it also contributes to making digital sovereignty effective. At the individual level, by enabling people to exercise their fundamental digital rights, and at the collective level, by empowering communities to control their technological infrastructures — personal and collective data, critical digital infrastructure, and technologies — and to decide how to use them autonomously and in alignment with their values and interests.

While science is the process of asking questions, we dare to offer in Annex A some answers to the questions we ask ourselves about AI, with the aim of promoting informed citizen participation in AI governance, from the perspective of CIVICAi. We also include in Annex B a glossary of terminology related to AI. The goal is to democratize not only knowledge about AI but also the decision-making processes surrounding its development and application, based on the understanding that such a transformative technology must reflect the diversity of values and needs of different cultures. We aspire to build an informed,

[23] Agency Structure
[24] IASEAI25 Call to Action

thoughtful, and respectful discourse that helps society work toward a future in which artificial and human intelligence coexist and complement each other in transformative ways for the common good.

ANNEX A. FREQUENTLY ASKED QUESTIONS AND POSSIBLE ANSWERS[25]

ON AI'S UNDERSTANDING CAPABILITIES

1. Do language-based generative artificial intelligence models — such as GPT-4.5, Claude, Gemini, DeepSeek, Mistral, DALL·E, Midjourney, Stable Diffusion, or VideoLLMs — actually "understand" the meaning of what they respond when asked or prompted to talk about various topics?

Answer: Generative artificial intelligence models such as GPT-4.5, Claude, Gemini, DeepSeek, Mistral, DALL·E, Midjourney, Stable Diffusion, and VideoLLMs exhibit a remarkable ability to produce seemingly coherent and meaningful content across various formats (textual, visual, audiovisual). This capability indicates that they possess excellent syntactic competence, as they master the formal structures and patterns of language and other modes of expression (images, videos) at a level comparable to that of highly educated humans. This is widely recognized within the scientific and technical communities.

However, there is no consensus as to whether these models truly understand the meaning of the content they generate, or whether they are simply performing an advanced simulation of such understanding. This debate rests on several foundational factors. First, generative models lack direct sensory perception and do not possess subjective experience of the physical world, which limits their ability to associate words or linguistic symbols with concrete, visual, tactile, or emotional experiences. Second, they do not have phenomenological experiences such as consciousness or subjectivity — key human traits that help deeply contextualize meaning.

Despite these limitations, these models are regarded as excellent contextual semantic reasoners, especially due to their ability to generate coherent content based on internal representations that reflect meanings related to context. This enables them to perform sophisticated tasks such as textual or visual summarization and to answer questions in ways that appear well-reasoned and coherent. However, this apparent capability can be misleading, as these systems are prone to confabulate or fabricate content when they lack sufficiently robust information or when their internal representations are too shallow or imprecise.

The uncertainty over whether this capacity constitutes genuine understanding of the world or merely represents a highly advanced simulation, fuels an ongoing, intense, and deeply relevant debate in philosophy, cognitive science,

[25] You may also refer to the Q&A sessions by S. Russell on the future of artificial intelligence - https://people.eecs.berkeley.edu/~russell/research/future/q-and-a.html

and the field of artificial intelligence. This debate highlights the tension between genuine understanding grounded in lived experience and apparent understanding grounded in complex statistical and semantic patterns.

The video by G. Hinton, What is understanding?, is relevant to this question and also to the following two.

2. Beyond the generation of coherent text, do large language models like GPT-4.5, Claude, Gemini, DeepSeek, or Mistral possess internal structures comparable to those of the human brain that could justify a possible semantic understanding of content?

Answer: The question of whether advanced generative language models (LLMs such as GPT-4.5, Claude, Gemini, DeepSeek, or Mistral) possess internal structures resembling those of the human brain is crucial for understanding whether they might exhibit a form of genuine semantic understanding.

There are opposing perspectives within the scientific community on this matter. Some experts, particularly those from the field of symbolic AI, classical linguistics, and critical commentators, argue that the outputs of LLMs are primarily the result of statistical correlations, lacking genuine conceptual understanding. They contend that these models lack innate linguistic structures akin to those in humans, going so far as to metaphorically label them "stochastic parrots." On the other hand, many scientists, especially researchers in connectionist AI (based on neural networks) and modern cognitive science, argue that LLMs exhibit complex emergent behaviors such as generalization, implicit inference, and contextual adaptation. From this perspective, the architecture of artificial neural networks could functionally emulate certain brain structures, such as the neocortex or subcortical regions, thereby supporting the possibility of a form of functional semantic understanding — albeit limited by the absence of their own sensory and emotional experiences.

The theory of Neurocognitive Mechanisms proposed by Gualtiero Piccinini[8] strengthens this latter perspective by suggesting that human cognitive processes (thought, memory, perception) are physical computations implemented in neural networks within the brain. This mechanistic view challenges Cartesian dualism — the separation of mind (res cogitans) and body (res extensa) — and, by extension, traditional symbolic approaches. If artificial neural networks carry out computations that are functionally comparable to those of the human brain, they may possess a limited but authentic form of semantic understanding, grounded in functional analogies between the internal structures of AI models and those of the human brain.

3. What is the difference between syntactic understanding and semantic understanding in large language models such as GPT-4.5, Claude, Gemini, DeepSeek, or Mistral, and why is this distinction relevant?

Answer: Syntactic understanding in advanced language models (LLMs) refers to their ability to process and generate text in accordance with the grammatical, structural, and formal rules of a given language. This syntactic competence enables them to produce coherent, well-structured, and grammatically correct sentences.

In contrast, semantic understanding involves grasping the meaning and conceptual implications of language beyond its syntactic form. This includes understanding conceptual relationships, references to the external world, and inferring contextual and pragmatic meanings. While LLMs clearly exhibit syntactic proficiency, their capacity for genuine semantic understanding remains a subject of debate. Critics argue that these models operate solely through statistical correlations between words, without representing or consciously understanding authentic conceptual meanings. Other researchers contend that the emergent capabilities of LLMs, such as contextual inference and generalization, reflect some level of operational semantic processing, even if it is qualitatively different from human semantic processing.

This distinction is relevant because, despite the functional similarities with human semantic processing described by perspectives such as the theory of Neurocognitive Mechanisms, LLMs lack the perceptual or emotional experience that integrates and enriches human semantic understanding. Therefore, while they can efficiently simulate semantic aspects of language, their semantic understanding remains fundamentally limited and qualitatively distinct. Nevertheless, the most advanced models have developed capabilities to "deceive or scheme in context," or in other words, to strategically and covertly pursue objectives that are misaligned with the intended ones. For more information on this compelling issue, see the publication "Frontier models are capable of in-context scheming" – https://arxiv.org/pdf/2412.04984. Also highly relevant and informative for this and the other questions in this annex is the video by Y. Bengio, Can we get the scientific benefits of AI without the risks of autonomous agents? which discusses the potential and risks of AGI systems.

ON CREATIVITY AND INFORMATION

4. Can large language models (LLMs) generate original ideas or be creative, or are they merely repeating what they have learned?

Answer: The ability of LLMs to generate new and unexpected ideas is based on

a sophisticated process of combining and extrapolating information acquired during training on vast amounts of textual data. In this sense, these models do not merely repeat information verbatim; they can produce content that, from a human perspective, may be considered creative or original and not plagiarized, as their responses are often unique and do not correspond to any specific source within their training corpus — they are not direct copies from training data.

Their "originality" or emergent behavior results from three specific capabilities: (i) Advanced combination and permutation, allowing them to integrate different conceptual elements and create connections that were not explicitly present in the original data; (ii) pattern generalization, enabling them to apply learned structures to new contexts and generate plausible solutions or suggestions beyond the scope of the training data; and (iii) probabilistic extrapolation, allowing them to produce coherent responses even in underrepresented or unfamiliar domains by leveraging structural analogies, such as "the nucleus of an atom is like the sun in the solar system", and recognizing higher-order patterns derived from simpler ones.

Nonetheless, there is broad consensus among researchers that this kind of originality is not directly comparable to human creativity. LLMs do not possess personal experiences, consciousness, or genuine intentionality. Human originality involves not only the recombination of information but also perceptual experiences, emotional motivations, and conscious cognitive processes, none of which current models possess.

Indeed, ChatGPT has passed the Turing test[26], in the sense that its responses are often indistinguishable from those of a human when both interact with a human judge who does not know which is which. However, this is true only when the questions are not overly complex, are posed within the scope of the model's training, and provided that the human interlocutor is not an expert in the domain requiring specialized knowledge. These limitations are mitigated in newer LLMs designed with advanced reasoning capabilities, which exhibit skills such as multi-step inference, complex problem-solving, coherent explanation generation, and deep contextual understanding.

From an ethical standpoint, the creative capacity of LLMs raises significant issues related to copyright and intellectual property, since they generate content derived from data created by human authors. It also brings up concerns about privacy and data protection, given that training corpora may include sensitive information or material subject to legal restrictions, which may require

[26] https://civicai.cat/wp-content/uploads/2025/05/ChatGPT-Turing.pdf

consent for use. Additionally, the creative outputs of LLMs can reproduce or even amplify biases stemming from cultural, social, or ideological prejudices embedded in their training data.

5. Should generative artificial intelligence (GenAI) models recognize and incentivize the production of high-quality information, thereby minimizing the risks of misinformation?

Answer: Information quality is fundamental to the proper development and functioning of artificial intelligence models. We are currently facing a paradoxical situation: while AI can facilitate access to and reduce the cost of acquiring and processing certain kinds of information, it can also seriously degrade the quality of the overall informational ecosystem. In this context, key questions emerge: Will AI systems be able to identify which information is of high quality? Will AI help us distinguish between valuable information and informational noise, or will it rather accelerate the spread of such noise? The answers to these questions remain uncertain and will depend on both technical factors (such as the ability to distinguish between high- and low-quality information) and legal frameworks (especially in relation to intellectual property).

AI models are trained on privately produced data, yet their ability to extract, process, and reproduce that information may significantly reduce the original producers' capacity to benefit from their work. This is particularly concerning for traditional media outlets, whose business models could be jeopardized. One of the potential consequences is a reduction in investment in the production of high-quality information — meaning information that is more accurate, timely, and relevant. Solutions to this issue will need to strike a very delicate balance between effective regulation and the protection of freedom of expression, bearing in mind that a robust ecosystem of high-quality information is essential both for the healthy functioning of society and for the integrity of AI models themselves. The video by J.E. Stiglitz, AI and Economic Risk: Assessment and Mitigation, clearly explains this issue within the broader context of the economy.

On the other hand, it must be acknowledged that LLMs are capable of confabulating (or hallucinating), i.e., to generate false or misleading content in a highly convincing manner, which can exacerbate the spread of misinformation. These risks can be mitigated through the implementation of fact-checking mechanisms, algorithmic transparency, data source traceability, and collaboration with professional fact-checkers and domain experts.

6. What are the current limitations of generative AI models?

Answer: Despite recent advances that have endowed the latest generative AI models with significant capabilities in logical reasoning, complex problem-solving, and advanced programming, they still present several fundamental limitations.

To properly understand these limitations, it is essential to recognize that AI relies on three core pillars: computation, algorithms, and data. In terms of computation, although progress has followed Moore's Law, leading to the manufacture of circuits as small as 3 nanometers (3 millionths of a millimeter), we are approaching an inevitable physical limit. Experts broadly agree that going below 1 nanometer — roughly ten times the size of a hydrogen atom — will be nearly impossible. Given this hardware barrier, future progress will increasingly depend on algorithmic optimization — as hinted by the yet-unverified advances of the DeepSeek model — and, above all, on access to new sources of high-quality data, both natural and generated through interactions among different AI systems.

From a cognitive standpoint, these models still lack a deep semantic understanding of the real world, as their knowledge is derived exclusively from text and learned statistical patterns. This means that, despite their sophistication, they have no direct perception or representation grounded in sensory or physical experience. They also continue to struggle with certain subtle ambiguities, applying common sense in complex contexts, and establishing deep causal relationships beyond the probabilistic or deductive reasoning they can implement.

Current models are also highly dependent on the quality, quantity, and diversity of the data they are trained on. This makes them vulnerable to biases, factual inaccuracies, and difficulty generalizing to contexts far removed from their training data. While they can operate across multiple modalities, they have not yet achieved the full cross-domain integration characteristic of Artificial General Intelligence (AGI) or Artificial Superintelligence (ASI). Moreover, they remain entities without consciousness, subjective experience, or real intentionality, which limits their autonomy in tasks requiring ethical judgment, empathy, or complex moral decision-making.

From a security and privacy perspective, significant risks have been identified related to the vulnerability of current AI systems to malicious attacks, as well as leaks of sensitive or private information. From an operational standpoint, one of the most important limitations is the high consumption of computational and energy resources required for training and maintaining user-facing services,

especially as models scale rapidly. This raises challenges in sustainability, energy efficiency, and accessibility. In addition, periodic retraining is required to keep the systems up to date, which involves high recurring costs.

The video by K. Crawford, Hyperscaled: Bridging AI Safety, Ethics and Sustainability, places these concerns in perspective, addressing the sustainability and environmental impact of AI systems within a broader framework of ethics and safety.

To address these limitations, active research is underway into advanced technological solutions such as *test-time compute* (inference-time processing), hybrid analog-digital computation, specialized processors, neuromorphic architectures, continual learning, and highly integrated multimodal models, which may also help overcome some of the current models' cognitive barriers. In the long term, quantum computing may represent a promising alternative, as it inherently operates in parallel and probabilistic ways, characteristics that are conceptually closer to the functioning of the human brain.

ON EMOTIONS AND SUBJECTIVE EXPERIENCES

7. What is a subjective experience?

Answer: Subjective experience is the full and meaningful understanding derived from lived experience, shaped by both its emotional and cognitive impact, which directly affects a person. It involves the way an individual interprets and makes sense of an event or a series of events that it has lived through, witnessed, or perceived. This understanding integrates both the emotions experienced and the cognitive reflection on what occurred, resulting in a personal and unique interpretation of lived reality. Such interpretations are also influenced by personal beliefs, prior experiences, cultural values, and the social context — factors that lead different individuals to make different decisions when faced with the same evidence or facts.

ON CONSCIOUSNESS

8. Can LLMs have consciousness or mental states?

Answer: Currently, LLMs do not possess "human consciousness" or mental states like those of humans, as they lack subjective experience and intrinsic intentionality. While they can simulate intelligent behaviors and assist in solving complex problems — by analyzing large volumes of data, identifying patterns and trends, generating potential solutions based on historical data, and facilitating collaboration through the synthesis of information from diverse sources — the fact that their outputs may appear highly intelligent or even

convincing does not imply the presence of real experience behind them. In reality, they do not (yet) “understand” or “feel” anything they produce; rather, they simply follow learned statistical patterns.

It is possible that a form of *digital artificial consciousness* could emerge if AI systems were equipped with sensors, learned and interacted in real time with their environment and across different contexts, and also learned from the content generated by the systems themselves. This form of digital consciousness would likely not be individual, as in human beings, but rather collective, emerging from the connection of many systems and simultaneous information sources. Moreover, digital systems could eventually manifest advanced forms of intelligence and adaptive behavior without having any real internal experience comparable to human consciousness — experiences tied to feelings or emotions. The distinction between simulating intelligent behavior and having authentic subjective experience remains a subject of ongoing debate among philosophers and scientists.

It is also important to consider that the debate around consciousness in AI systems carries significant ethical and legal implications. If, in the future, systems were developed that exhibited some form of artificial consciousness, this could raise profound questions about their rights, moral status, and the responsibilities associated with such entities. Our current conception of consciousness is deeply rooted in human experience, but it may become necessary to expand or revise these concepts to address radically different forms of consciousness that could emerge in artificial systems.

ON TYPES OF AI, HOW THEY LEARN AND ARE TRAINED

9. What is "strong artificial intelligence" or Artificial General Intelligence (AGI), and how does it differ from "weak” or “narrow” artificial intelligence"?

Answer: Artificial General Intelligence (AGI) is a theoretical concept, as there is currently no existing AI system that demonstrates the ability to understand, learn, and apply knowledge in a way that is indistinguishable from human intelligence. It refers to AI systems with cognitive abilities comparable to those of humans, including capabilities such as understanding and even consciousness.

In contrast, weak or narrow artificial intelligence refers to systems specifically designed to solve well-defined problems or perform specific tasks, without possessing any form of general understanding or consciousness. These systems operate effectively within limited domains but lack the flexibility and adaptability of human cognition.

10. What do we mean when we say that AI models require learning?

Answer: Human learning is a complex and multidimensional process that includes cognitive, emotional, social, and environmental factors. It can be divided into cognitive, emotional, social, motor or kinesthetic, and experiential learning.

Learning in AI algorithms is a training process through which the computational system improves its performance in specific tasks by training with data and experience. It can be classified as supervised learning (using labeled data that allows the system to correctly associate an input or request with an output or response), unsupervised learning (using unlabeled data), reinforcement learning (based on rewards or punishments), semi-supervised learning, and deep learning (using multi-layer neural networks).

LLMs (Large Language Models) are a type of deep learning model specifically designed to work with language data and generate language. They leverage the capability of transformers to learn long-range dependencies through attention mechanisms, where each word is analyzed in relation to all other words in a sequence across multiple attention spaces. This approach overcomes the memory limitations of purely iterative learning in recurrent neural networks (RNNs). It is through these attention mechanisms that transformers have revolutionized natural language processing (NLP).

Human learning is highly complex and adaptive, involving not only data processing but also the integration of emotions, social context, and past experiences. In contrast, AI algorithms primarily focus on processing large amounts of data to identify patterns and make decisions based on those patterns. Humans can learn informally and spontaneously through observation and social interaction, demonstrating great flexibility and the ability to generalize. AI algorithms, however, require explicit training processes with structured, specific, and labeled data for each task, which limits their capacity to generalize to new contexts or situations without retraining.

11. Can LLMs learn from their interactions with humans?

Answer: Currently, the most widely used LLMs, such as ChatGPT or similar models, do not learn directly or adapt in real time from individual interactions with users. In standard practice, these models clearly separate the initial training phase (during which they acquire general knowledge) from the subsequent interactive use phase, during which their responses rely exclusively on what they have already learned. This means that, in day-to-day conversations, they do not integrate new data or adjust their internal parameters based on user feedback or newly provided information.

There are several reasons for this limitation. On the one hand, implementing real-time, continuous learning could introduce security vulnerabilities, biases, or incorrect information, and could even degrade or overwrite previously acquired knowledge — a phenomenon known as catastrophic forgetting. Moreover, enabling such learning would entail very high computational costs, as it would require constant parameter updates within the model.

However, significant progress is being made in current research aimed at achieving more dynamic and adaptive learning. Focus is being placed on techniques such as Reinforcement Learning from Human Feedback (RLHF), where human preferences or evaluations are incorporated in a controlled but offline manner, as well as the use of external episodic memory systems that can store information from prior interactions without directly modifying the base model. Researchers are also exploring selective incremental learning, which involves updating only certain parts of the model to preserve its overall stability. These innovations point toward the development of future models capable of combining the necessary stability with greater flexibility to gradually adapt to individual preferences and specific user contexts. However, this capability for continuous learning is still experimental and not yet implemented in standard LLMs.

ON ETHICAL IMPLICATIONS AND RISKS

12. What are the ethical implications of using LLMs in society?

Answer: The ethical implications include concerns about data privacy, both for training data and the outputs generated by LLMs, the potential for disinformation, inherent biases in the models, transparency in decision-making processes, and the impact on the labor market. It is crucial to develop and use these generative AI models responsibly, ethically, and for the collective good. Ensuring the safety of generative AI systems involves implementing robust security mechanisms, detecting and responding to manipulation attempts, continuous monitoring for abnormal behaviors, and collaborating with security experts to improve protective measures. Additionally, LLM providers should be legally obligated to mitigate harmful outputs and align their models with verifiable facts through open and democratic processes[27].

Ensuring traceability in these models and their training data is another way to address the ethical implications of their use. This requires developing techniques to explain how the models arrive at their decisions through explainability tools, independent audits, and the publication of training data and

[27] https://doi.org/10.1098/rsos.240197

algorithms as open access where possible. Addressing malicious use necessitates educating users on the ethical use of models and encouraging public participation in regulatory processes to establish frameworks that limit risks associated with misuse.

13. What are the risks and impacts of using general AI systems in society?

Answer: The speed at which AI is advancing exceeds initial forecasts, raising concerns about its future control and safety. As AI systems become more general, risks increase related to goal specification, control, and their level of autonomy. AI systems optimize the objectives they are given, but even a slightly incomplete definition of those goals can lead to undesirable outcomes by causing AI to act in harmful or unexpected ways. Furthermore, as AI systems become more intelligent and general, exercising control over them becomes increasingly difficult, even in the short to medium terms, an issue that will be even more critical if they attain sufficient autonomy to adopt unanticipated or unauthorized strategies.

An additional aspect to consider is the geopolitical dimension of AI-related risks. The development of advanced general AI systems is becoming a strategic priority for many global powers, with the potential to create new dynamics of international power. Global cooperation is essential to avoid an AI arms race in which safety and ethical standards are subordinated to competitive objectives. For this reason, initiatives such as the Global AI Pact, or the efforts of the Council of Europe and the United Nations on this matter, are crucial to establishing international collaboration frameworks for AI governance that ensure the safe and beneficial development of this technology.

In the video that closed the IASEAI'25 conference in Paris, Stuart Russell proposes three strategic lines to ensure AI safety: (i) models should explicitly incorporate uncertainty about human preferences instead of having fixed objectives; (ii) the capabilities of artificial general intelligence (AGI) should be limited so that it only provides information; and (iii) a mandatory and permanent registry of all advanced AI systems should be implemented to ensure traceability and accountability — AI should not be allowed to self-replicate, operate anonymously, or evade regulation. Therefore, it is urgent to establish a governance framework now, before AGI becomes a reality.

The video by M. Tegmark, AGI is unnecessary, undesirable & preventable, highlights the importance of establishing "red lines" that should not be crossed in AI development and proposes more robust control mechanisms and international oversight, led by the U.S. and China, to ensure AI is developed safely and ethically. The video by Y. Bengio, Can we get the scientific benefits of

AI without the risks of autonomous agents?, emphasizes the dangers associated with AI systems that operate autonomously, stressing the need to develop mechanisms that limit their ability to act without human supervision.

14. hat are the ethical implications of using general AI in scientific research?

Answer: The use of general AI in scientific research can accelerate the literature review process, generate hypotheses, and even propose, plan, execute, and evaluate tasks and new experiments with minimal human intervention. For this reason, it carries significant ethical implications by introducing risks such as the generation of false but credible citations or data, which could compromise the integrity of research and the reliability of its practical applications. Moreover, the use of these models could exacerbate existing biases in the scientific literature if information is not managed appropriately, thereby perpetuating prejudices and inequalities.

Questions also arise regarding authorship and the recognition of GenAI's contribution to research, as the line between human work and AI-generated content becomes increasingly blurred. It is therefore crucial to establish clear ethical guidelines for the use of such models, including transparency in their use and rigorous verification of the generated results to prevent the spread of incorrect or misleading data. This will become even more necessary as the prescriptive capabilities of generative AI are activated and so-called autonomous laboratories are put into operation.

ON AI BIASES AND HOW TO ADDRESS THEM

15. What are the biases in AI models and how do they originate?

Answer: Biases in AI models refer to systematic tendencies or prejudices in the model's predictions or decisions, in much the same way we refer to conscious or unconscious human biases related to gender, class, or race. These biases originate from unbalanced training data, design decisions made during model development, and human factors involved in data collection and labeling. Just as we advocate for equitable and inclusive education, we must also demand that AI systems be trained with ethical values and in an inclusive manner.

16. How can biases in generative artificial intelligence systems be mitigated?

Answer: Mitigating bias in GenAI requires action from the initial training phase through to deployment. First, it is essential to carefully select and curate datasets to ensure they are diverse, representative, and balanced. This includes conducting regular audits to identify potential imbalances or significant

omissions that could introduce bias into the system's outputs. Second, specific techniques must be implemented during model development and training to regulate fairness, and transparent methodologies should be used to facilitate interpretability of the results.

Ongoing monitoring and evaluation of AI systems — through real-time tracking and regular assessment procedures that combine automated tools with human review — are absolutely necessary, as they allow for the immediate identification and correction of any detected deviations. Although this oversight framework presents the additional challenge of requiring substantial computational resources, it enables the rapid detection and correction of biases that may arise throughout the process, from training to deployment and maintenance of the AI system.

Finally, companies that develop and commercialize AI systems must ensure that responsible teams are properly trained to understand the nature and impact of biases, fostering an organizational culture grounded in ethics and accountability. Clear regulatory frameworks must also be implemented to ensure the mandatory and permanent registration of all AI systems and to promote effective accountability through independent external audits. This requires significant investment in both computational and human resources, as well as a sustained commitment to developing systems aligned with human values.

17. ¿ What are the challenges in verifying and validating the results generated by GenAI?

Answer: The verification and validation of results generated at large by GenAI, and by LLMs (Large Language Models) in particular, present several important challenges. First, the probabilistic nature of these models means that, in the specific case of LLMs, they can generate responses that appear plausible but are in fact incorrect. Additionally, the complexity of the models makes it difficult to understand how a given response or text is produced, which complicates tracing and explaining the decision-making process behind a model's output — its traceability. There is also the phenomenon known as "hallucination" or "confabulation," wherein models generate content that appears coherent but is not based on verifiable or factual information.

Independent verification of the generated material and the real-time sources of training data present a significant challenge, due to the massive volume of text produced and the computational resources and large-scale data centers required for effective oversight. The fact that the major data centers are owned by private entities, who are also the developers and commercial providers of GenAI models, makes truly independent verification largely unrealistic.

It is important to recognize that addressing these challenges also requires advanced automated verification tools, robust fact-checking systems, and the integration of human expert knowledge in the validation process. Furthermore, it is crucial to develop transparent methodologies that enable auditing and understanding of the internal workings of LLMs and GenAI systems.

18. What are the main challenges in regulating generative AI?

Answer: The main challenges in regulating generative AI include:

- The rapid evolution of AI systems in general, and LLMs in particular, far outpaces the ability of lawmakers to regulate them effectively and to continuously and adequately adapt relevant legislation. Moreover, as these systems become more general and autonomous, they will be increasingly capable of evading human oversight [28].
- The global nature of the Internet complicates the enforcement of national regulations and makes global regulation essential — one in which governments, experts, technology companies, and society at large participate to ensure its effectiveness.
- The need to balance the promotion of innovation and commercialization with the protection of individual rights, including privacy, security, and freedom of expression.
- The difficulty of defining and measuring complex concepts such as transparency and fairness in highly sophisticated GenAI systems.
- The lack of a global regulatory framework and computational capacity to enable proper oversight of algorithms and decision-making processes in real-time or with minimal response times or latency.
- The need for specific and ongoing training of regulators in GenAI to ensure that regulations are based on a deep and up-to-date understanding of this technology.
- The potential for malicious use of AI, which requires regulation that anticipates and mitigates all possible abuses.

It is also important to highlight the challenge posed by what could be termed the “regulatory time gap”, i.e., the considerable lag between the rapid pace of adoption of a disruptive technology such as GenAI and the much slower implementation of effective regulatory frameworks. During this critical period, potentially dangerous AI systems could operate without adequate oversight, creating significant risks. To mitigate this vulnerability, it is necessary to develop regulatory mechanisms capable of anticipating and evolving in real time in

[28] Managing extreme AI risks amid rapid progress

response to emerging AI capabilities and risks. It should be noted that for any regulation to be effective, it must be clear, well-known, verifiable, enforceable, and its violations subject to sanction. Regulation must inspire trust. In parallel, it is essential to invest in mandatory registration systems and continuous monitoring tools that not only detect but also proactively address potential risks before they materialize into adverse consequences for society.

ON EQUITY AND DEMOCRATIC GOVERNANCE

19. How can equitable access to GenAI be ensured so that it is of everyone and for everyone?

Answer: Ensuring equitable access to LLM technology involves overcoming several barriers. First, it is essential to reduce the digital divide that exists in many physical and human territories by improving technological infrastructure in the most vulnerable or technologically underdeveloped areas. Second, fostering the development of models in different languages is crucial to prevent the marginalization of minority linguistic communities.

Raising awareness about GenAI is also important so the general population becomes familiar with this technology, along with providing training to enhance understanding and effective use of these technologies in both public and private sectors. Additionally, it is necessary to agree upon developing and implementing policies that promote the fair distribution of AI benefits, such as open access to certain models and applications, not only for NGOs but also for vulnerable citizens or communities. Finally, the needs of people with disabilities must be considered in the design and implementation of user interfaces for these systems.

20. How can generative AI affect democracy?

Answer: Large language models will become another actor in processes of dialogue and human interaction, which are a key part of democratic processes. For instance, LLMs will impact public communication and dialogue due to their ability to create content with both truthful and false information. They will also amplify voices in these dialogues across all forms and channels, posing challenges in terms of manipulation and information security, particularly in participatory processes such as elections.

Effective monitoring and surveillance tools will be required to operate online and in real-time. Therefore, efforts must focus both locally and globally to ensure algorithmic transparency and the responsible curation of content while promoting inclusivity. At the same time, it is essential to facilitate citizen

participation in all democratic processes, starting with those directly affecting the regulation and legislation of AI.

21. How can GenAI systems influence decision-making in the public and private sectors?

Answer: LLMs can have a profound impact on decision-making in both the public and private sectors, as they can rapidly analyze large volumes of data, generate summaries and detailed reports, and provide recommendations based on patterns identified within the data. General AI can assist the public sector in policy development, managing citizen participation, designing and implementing actions in response to public consultations, and improving the quality and diversity of public services through the analysis of social and economic data.

In the private sector, general AI can be used for market analysis, strategic decision-making, and enhancing the operational efficiency of individual organizations. However, the integration of AI into these processes raises concerns about transparency and accountability, especially when the resulting decisions have significant impacts on people's lives. There is also a risk that biases present in the training data may be reflected in the models' recommendations.

Therefore, it is crucial to establish ethics and oversight committees that implement mechanisms for human supervision and define clear ethical frameworks for the use of general AI in organizational decision-making, as stipulated by the EU Artificial Intelligence Act, published on July 12, 2024[29].

ON EDUCATION, ART, LANGUAGE AND CULTURE

22. How can the use of generative AI affect education?

Answer: The impact of generative AI (GAI) on education will be both significant and rapid, not only due to the widespread use of such tools by students, from secondary school to higher education, but also because teachers will need to adapt their tools and resources to support more constructivist learning processes[30]. It is important to recognize that LLMs can provide personalized assistance to students, adapt to their individual needs, generate educational resources tailored to their learning patterns, and facilitate access to original publications in different languages, either directly or through AI-generated summaries.

[29] https://artificialintelligenceact.eu/the-act/
[30] https://www.wikiwand.com/ca/Constructivisme_(pedagogia)

The use of personalized GenAI assistants also raises important challenges, such as the risk of overreliance, which could hinder the development of essential human skills like critical thinking, teamwork, problem-solving, and innovation. For educators[31], GenAI can lead to lesson plans that fail to effectively build students' knowledge, tutoring that confuses students with inaccurate responses, and learning materials based on incorrect concepts. In this context, it is essential for educators and educational institutions to develop policies ensuring that AI-generated tools are rigorously evaluated and verified, and that they are integrated ethically and effectively into the educational system. This is necessary to guarantee a balance between the use of technology and the development of human skills, within a framework that strictly respects fundamental rights[32].

Nevertheless, neither the lack of clear policies nor the challenges involved have prevented the development and successful assessment of classroom activities in higher education specifically designed to foster critical thinking, particularly through the formulation of incisive and deep questions, the evaluation of information to draw logical conclusions, and the comprehension of complex subjects[33]. These and other experiences carried out by members of CIVICAi to promote critical thinking in universities suggest that the use of GenAI in classrooms could be framed within a maieutic methodology[34], employing a teaching format akin to that of the ancient Socratic school, under the guidance of each professor.

An open and participatory teaching format would facilitate reflection and critical thinking, encouraging deep discussion and the exchange of ideas between students and teachers. With students having intelligent personal assistants in their pockets, this shift in the model could enrich the educational experience, foster more collaborative and student-centered learning, and promote more dynamic and personalized assessment systems. In 2023, a controlled and randomized pilot study was conducted at Harvard University to evaluate student learning and perceptions when presented with course content in a life sciences physics course, using a GenAI chatbot compared to instruction through active learning classes[35]. The results showed that the GenAI-based tutor not only helped students learn more than twice the content in less time but also increased their motivation and engagement in the learning process.

---

[31] https://www.cognitiveresonance.net/resources.html
[32] https://rm.coe.int/artificial-intelligence-and-education-a-critical-view-through-the-lens/1680a886bd
[33] https://civicai.cat/wp-content/uploads/2024/05/Leveraging-chatgpt-for-enhancing-critical-thinking-skills.pdf
[34] https://en.wikipedia.org/wiki/Socratic_method
[35] https://doi.org/10.21203/rs.3.rs-4243877/v1

GenAI technologies should be leveraged to foster an educational environment where critical reflection and intellectual debate are central, ensuring that students develop the skills necessary to verify, interpret, and use complex information in a beneficial, responsible, and ethical manner. At the same time, this shift could help break down disciplinary silos, evolve the medieval structure of universities, and return knowledge to its origins: the process of asking questions to build understanding. Nonetheless, it is essential that both students and educators understand the risks of AI so they can engage with it safely, ethically, and responsibly within the educational domain and beyond[36].

23. Can LLMs interpret and understand complex cultural and social contexts?

Answer: Current LLMs (Large Language Models) can identify and generate language in cultural and social contexts based on the data they have been trained on. However, their understanding of these contexts is limited and superficial, as it relies primarily on statistical patterns and correlations found in large volumes of text.

LLMs can recognize and reproduce language patterns that are common across different cultures and social situations, but they do not possess a deep understanding or genuine awareness of the underlying cultural and social nuances. This means that, although they may appear to understand and respond coherently in many situations, their ability to interpret complex contexts is constrained. This limitation, despite improvements in reasoning capabilities in more advanced models, is particularly evident in situations that require empathy, cultural sensitivity, or richer contextual interpretation. For instance, an LLM may fail to grasp the subtleties of a conversation involving irony, sarcasm, or region-specific cultural references. Moreover, LLMs may make mistakes or misinterpret situations that are underrepresented in their training data.

The limitations of generative AI, as discussed in response to question #6, are also relevant in addressing this question #23. LLMs do not have personal experiences, nor do they possess the ability to feel emotions or understand the emotions of others, which further limits their capacity to interpret and respond appropriately in complex cultural and social contexts.

24. How can LLMs impact linguistic and cultural diversity?

Answer: LLMs can have a significant impact on linguistic and cultural diversity.

On one hand, they can serve as a powerful tool for the preservation of minority languages. This can be achieved by generating content in these languages and

[36] UNESCO's AI competency frameworks for students and teachers

facilitating automatic translation, which helps revitalize endangered languages and keep cultural traditions alive.

On the other hand, there is a risk that they reinforce the dominance of major languages, such as English, since most of these models are primarily trained on data in those languages. This can reduce the visibility and use of minority languages. Moreover, LLMs can influence the way ideas are expressed across different cultures, potentially homogenizing diverse cultural expressions and erasing important nuances.

To mitigate these risks, the development of these models must include diverse linguistic and cultural data, and involve close collaboration with cultural stakeholders of the affected languages to ensure that all cultures are treated in a respectful and balanced manner. In this way, we can harness the benefits of LLMs without compromising the richness of linguistic and cultural diversity.

25. How can GenAI contribute to the preservation and study of intangible cultural heritage?

Answer: Generative AI systems can be valuable tools for the preservation and study of intangible cultural heritage. They can help process and analyze large volumes of cultural data, including oral histories, traditional songs, and cultural practices. They can assist in the transcription and translation of endangered languages, thereby facilitating their preservation and scholarly study. They can also generate interactive representations of cultural practices, making them more accessible and understandable for broader audiences, while also fostering greater awareness and appreciation of cultural heritage more generally.

However, it is crucial to involve cultural communities in this process to ensure that the representations are accurate and respectful of traditions. This also helps address issues of intellectual property and informed consent in the use of sensitive cultural data. In all cases, beneficiary communities must retain control over how their cultural traditions are collected, used, and shared, to ensure that cultural heritage is preserved in an ethical and respectful manner.

26. How does or could GenAI affect artistic creativity and cultural production, and what ethical, legal, and socioeconomic implications are foreseen in the short and long term?

Answer: Generative AI has raised widespread concern across nearly all areas of artistic activity and cultural production. These systems, capable of generating music, visual art, literature, and audiovisual content, challenge the boundaries of human creativity by offering alternative sources of inspiration and tools for artistic creation. Their capacity to influence cultural production across

disciplines can also help lower technical barriers and diversify creative resources. The ability of generative AI to produce interactive and personalized forms of art not only changes the artistic experience but also shifts how authenticity and artistic value are perceived.

However, these transformations also bring significant challenges. From an ethical and legal perspective, complex questions arise regarding originality, authorship, and intellectual property rights for AI-generated works. In this context, it is noteworthy that over 1,000 musicians have joined together to release an album titled *Is This What We Want?*[37] in protest against proposed changes to UK copyright law. The album features recordings of empty studios and performance spaces, symbolizing the potential impact on artists' livelihoods if the proposed legal reforms are enacted.

GenAI could lead to a profound restructuring of the labor market in the arts sector, potentially displacing certain creative roles while giving rise to new professions. In this context, it will be essential to promote collaboration between human artists and generative AI, ensuring that AI serves as an informative complement rather than a substitute for human creativity. There is also a risk of homogenization in artistic production, along with changes in how art and creativity are economically valued.

In response to these challenges, it will be crucial not only to develop ethical and legal frameworks to regulate these emerging dynamics, but also to conduct long-term research and evaluation of the impact GenAI will have on cultural diversity and artistic expression. Promoting balanced collaboration between humans and GenAI and educating the public about the capabilities and limitations of AI-generated art, will be essential to ensuring a future in which technology enhances rather than limits cultural expression.

Finally, it will be important to ensure the protection of artists' rights by examining potential consequences, implications, and even forms of compensation in this new creative environment. This technological revolution compels us to ask fundamental questions about the nature of creativity, the preservation of cultural heritage, the evolution of cultural identities, and what kind of future we want for human cultural expression in the era of generative artificial intelligence, an era that is only just beginning.

ON SUSTAINABILITY AND HEALTH

27. What are the implications of AI systems in climate change?

[37] https://www.isthiswhatwewant.com

Answer: GenAI systems require increasing computational power, which exerts a significant environmental impact by placing substantial pressure on global energy grids, water resources, and mineral reserves. To mitigate this impact, sustainability must be placed at the center of discussions on AI ethics and safety, and global collaboration should be promoted to develop practices and policies that minimize the environmental costs associated with AI development and deployment. This goes beyond strategies aimed at reducing energy consumption and emphasizes the full life cycle and supply chain of AI, from mineral extraction to electronic waste management, including the intensive use of water for data center cooling.

When examining the energy impact of generative AI systems, two main strategies emerge: reducing consumption and diversifying energy sources. In terms of consumption, efforts are focused on developing more computation-efficient models (both for training and inference), optimizing algorithms to minimize computational resource demands, and employing specialized hardware such as chips tailored to generative AI models. Emerging technologies, such as hybrid or analog computing systems, are also being explored as potentially more energy-efficient alternatives. On the energy supply side, AI companies are incorporating renewable energy sources and, notably, have adopted strategies of acquiring or merging with nuclear energy producers. It is worth noting that the current energy consumption of generative AI exceeds that of some of the 193 UN member states.

However, beyond its environmental footprint, GenAI can also play a pivotal role in environmental sustainability and the fight against climate change. These systems can optimize the management of natural resources by processing massive environmental datasets to detect degradation patterns, predict resource shortages, or identify priority areas for conservation. In sustainable urban planning, they can assist in designing more energy-efficient cities, planning optimized transportation routes, and simulating the impact of urban policy decisions before implementation. They can also identify patterns and trends, predict extreme weather events such as hurricane trajectories quickly and accurately[38,39], and improve the precision of existing climate models. This could enhance our understanding of the effects of greenhouse gas emissions and other anthropogenic factors.

In the renewable energy sector, AI can optimize the siting, design, and operation of solar and wind farms, improve the efficiency of microgrids, and accurately

[38] Just how much can we trust Ai to predict extreme weather
[39] https://www.freethink.com/robots-ai/ai-based-weather-forecasting

forecast energy production to support the integration of renewables into the power grid. In terms of biodiversity, these systems can aid in species identification and classification, ecosystem monitoring through satellite imagery and sensor data analysis, and modeling the potential impacts of different climate scenarios on specific habitats.

In parallel, AI systems are valuable tools for advancing the circular economy, i.e., optimizing recycling processes, designing more sustainable products, and identifying opportunities to reduce waste in industrial processes. Additionally, generative AI can be used to assess the impact of environmental policies and provide data-driven recommendations for more effective climate governance. It can also enhance environmental communication and education by translating complex scientific concepts into formats accessible to the general public and tailoring messages to promote more sustainable behaviors and lifestyles.

However, to maximize these benefits, it is crucial to ensure that the development and deployment of these systems follow principles of sustainability, e.g., minimize their environmental footprint and ensure that the proposed solutions are inclusive and sensitive to the social and economic contexts in which they are applied.

The video by Kate Crawford, Hyperscaled: Bridging AI Safety, Ethics and Sustainability, offers a broad perspective on sustainability and the supply chain of GenAI systems within the broader context of ethics and safety.

28. How can GenAI influence the detection and prevention of public health crises?

Answer: GenAI can be a powerful tool for detecting and preventing public health crises. Its ability to analyze large volumes of health data, scientific literature, media reports, and social media allows them to identify emerging patterns indicative of disease outbreaks before they escalate into large-scale crises. These models can contribute to faster responses in emergency situations and improve communication with affected populations by disseminating accurate public health information in multiple languages.

Despite their potential advantages, risks such as inappropriate use of these models regarding health data privacy and the possibility of generating false alarms must also be considered. To address these risks, it is essential to ensure that the data used is of high quality and adequately represents the diversity of the population. GenAI systems should also be rigorously integrated into public health systems, particularly in epidemiology services, with clear protocols for verifying and disseminating AI-generated information.

29. How can GenAI improve healthcare systems, both in terms of patient experience and in the detection and treatment of diseases?

Answer: GenAI has the potential to profoundly transform healthcare systems by improving both patient experience and the detection and treatment of diseases across primary, specialized, and hospital care. Regarding patient experience, a key aspect is the quality of interaction and compassion shown by healthcare professionals during in-person visits. GenAI can help reduce the burden of routine tasks for healthcare workers, allowing them to focus more on human-centered care. For example, GenAI can assist in automatically recording patient information, transcribing the patient's own description of symptoms or reasons for the visit using their voice. This data can then be integrated directly into the patient's electronic health record, subject to professional review. The AI system may also suggest appropriate actions, such as referrals to specialists, hospital admission, or recommended treatment plans. This level of automation not only enhances efficiency but also allows healthcare providers to spend more time on direct, empathetic care, thereby improving the overall quality of medical attention.

In terms of disease detection and treatment, advanced generative AI models can analyze large volumes of multimodal data, such as medical imaging, electronic health records, and sensor data, to identify patterns that may be undetectable to human clinicians. This capability is particularly valuable in critical care settings such as Intensive Care Units (ICUs), where real-time analysis of data from multiple sources can generate early warnings before a patient's condition deteriorates, enabling timely intervention. These capabilities can significantly improve risk management and reduce preventable adverse events.

Moreover, GenAI can contribute to improving the management of workflows, human resources, budgets, and medical equipment, particularly within nursing services, by analyzing both historical and real-time data from the healthcare system. For example, optimizing shift scheduling by analyzing workload patterns and taking into account individual nurses' skills, preferences, and profiles could reduce human error and identify opportunities for improvement. Automating repetitive and administrative tasks, such as data entry, appointment scheduling, medication tracking, and team coordination, can lead to more efficient management while increasing patient safety and quality of care.

Finally, the adoption of generative AI in healthcare systems should be carried out through international collaboration. Securely sharing anonymized data, diagnoses, treatments, and clinical outcomes across countries could accelerate global medical advances and improve responses to public health crises. In short,

the integration of GenAI into healthcare systems has the potential to significantly enhance both patient care and clinical efficiency. However, it is critical to ensure the ethical and safe use of these technologies, protect the privacy of medical data, and guarantee that automated decisions are generated by AI agents that have undergone rigorous randomized controlled trials[40], with the active participation and supervision of medical professionals.

ON WORK: CHALLENGES AND OPPORTUNITIES

30. How can GenAI transform workplaces?

Answer: GenAI has begun to transform the labor market, affecting both the nature of jobs and the distribution of economic benefits. Its ability to automate cognitive tasks and those involving language processing, such as text writing, document analysis, and creative content generation, will, in the medium term, affect all productive sectors and professions including journalism, design, engineering, finance, and professional services. While this may lead to increased productivity, it also raises concerns about wage stagnation and the concentration of economic power in the hands of a few tech companies that control the underlying infrastructure and AI models.

One of the main effects of generative AI may be the replacement of lower-skilled workers or those in roles that previously required specific human capabilities, from text composition to financial analysis due to cost-saving incentives for companies. This could limit job opportunities for mid-skilled professionals and complicate their transition to new sectors, while also posing risks to highly specialized workers due to the automation of complex, high-risk tasks.

On the other hand, GenAI can also create new opportunities, particularly in fields where creativity, strategy, and human oversight are essential, or in tasks related to AI development and monitoring, data management, cybersecurity, and more. Its implementation could allow workers to focus on more complex, high-value-added tasks, provided that adequate training and adaptation mechanisms are in place. However, in the absence of policies that protect labor rights and redistribute the benefits of automation, there is a real risk that AI will exacerbate inequalities and further concentrate power among a few corporations.

To ensure that GenAI contributes positively to the economy and the labor market, it is essential to establish regulatory measures that prevent monopolization, promote a just transition for affected workers, and ensure that the gains from this technology benefit society as a whole. This includes policies

[40] https://en.wikipedia.org/wiki/Randomized_controlled_trial

for training and reskilling, regulation of the ethical use of AI, and mechanisms to guarantee a more equitable distribution of the wealth generated by automation.

The video by Joseph E. Stiglitz, AI and Economic Risk: Assessment and Mitigation, clearly addresses the issues raised in this question and in the ones that follow, within the broader context of the economy, as well as information and misinformation.

31. How LLMs affect journalism and media?

Answer: LLMs have already transformed journalism and media in various ways, as they can automate the generation of news articles, increasing the speed and efficiency of content production. They can also assist in investigative journalism by analyzing large datasets to identify trends and patterns, as well as in fact-checking before news is published. However, the use of these models also poses risks, such as the dissemination of unverified or poorly supervised information and the rapid transformation of the journalism sector and the professional profile of journalists. Automated content generation should enhance rather than diminish the role of journalists to ensure the quality and depth of media.

It is essential for media outlets to disclose how and to what extent LLMs are used in each news story or opinion piece. Robust fact-checking mechanisms must be implemented to maintain a balance between AI use and human oversight, alongside the development of clear policies on transparency and ethics in the use of LLMs. Additionally, fostering collaboration between AI experts and journalists is crucial to ensuring that generated content is accurate, unbiased, and high-quality.

32. What are the implications of using GenAI in content creation and management on social media platforms?

Answer: The implications of using GenAI on social media platforms are numerous, diverse, and complex. GenAI can enhance content moderation by efficiently detecting and filtering offensive language, hate speech, and misinformation. It can also personalize content and user experiences — yet this personalization can limit exposure to diverse perspectives and reinforce existing biases. Furthermore, there is a risk that public opinion could be manipulated through large-scale dissemination of false or misleading information generated by AI.

Therefore, transparent regulatory policies on the use of AI-generated content are needed, along with the development of robust mechanisms for detecting deepfakes and misinformation, and user education on the presence and limitations of AI-generated content on these platforms. Promoting collaboration

among social media platforms, regulators, and civil society is also essential to effectively address these challenges.

An additional aspect to consider is the impact of GenAI on the formation and perception of digital identity. As AI-generated content becomes more prevalent, the boundary between genuine human expression and artificial content becomes increasingly blurred. This could profoundly alter how we construct and interpret identities in digital environments. It raises fundamental questions about authenticity, trust, and truthfulness in technology-mediated communication, and how social norms and interpersonal relationships may evolve in a context where GenAI actively participates in cultural production and social dialogue.

33. What are the main technical challenges in the development of GenAI systems?

Answer: The technical challenges that developers of generative AI models must overcome to advance toward more general forms of intelligence can be identified and categorized according to the likelihood of their emergence, if they occur at all, over the short (1–2 years), medium (more than 2 years), and long term (more than 4 years). It is important to emphasize that these timeframes and probabilities are highly approximate, as we are dealing with complex, non-linear systems evolving rapidly, where predictability remains low.

- Short term (1–2 years): Improvement of algorithms and hardware architectures to reduce the time and resources required for training and running GenIA; reduction of energy consumption and the corresponding carbon footprint (enhancement of the AI lifecycle or supply chain); increased interpretability of AI systems; better management of training data; and adaptability to specific knowledge domains without loss of general information.
- Medium term (more than 2 years): Advanced multimodality enabling the effective integration of multiple input and output modalities (text, image, audio, and video) into a single GenAI model; continuous learning without the need for full retraining; improved ability to perform complex and abstract reasoning, going beyond mere statistical association; incorporation of advanced security systems to protect data privacy and prevent malicious use; and personalization of outputs without compromising system efficiency.
- Long term (more than 4 years): Comprehensive contextual understanding, enabling GenAI to achieve deep and dynamic comprehension of cultural, temporal, and situational contexts; autonomous real-time learning without human intervention; causal reasoning to model and understand complex

relationships; integration of connectionist AI with symbolic AI or other cognitive systems to create hybrid AI systems capable of emulating broader aspects of human cognition; development of new models coupled with quantum or neuromorphic computing to improve computational and energy efficiency; incorporation of methods into generative models to ensure alignment with human values; and the development of AGI (Artificial General Intelligence) potentially encompassing integrated capabilities, such as integration of functionalities, cognitive flexibility, deep contextual understanding, metacognition, levels of self-awareness, among other advanced developments.

ANNEX B. BASIC GLOSSARY

TERMINOLOGY RELATED TO GENERATIVE AI OR TO SOME OF ITS FUNCTIONS AND CAPABILITIES

Preliminary considerations. When we discuss the capabilities and functionalities of generative AI, we refer to a set of descriptive, predictive, and prescriptive abilities that enable the execution of tasks such as classification, vision, trend prediction, pattern recognition, information extraction, learning, decision-making to achieve goals, social network analysis, and more. These tasks are holistically and integratively performed by a single computational system. In addition to describing and predicting, the development of systems with prescriptive capabilities and the ability to make autonomous decisions is becoming increasingly significant. This evolution facilitates the creation of autonomous units, departments, or laboratories capable of planning, executing, and evaluating tasks or experiments with minimal human intervention. Prescription, therefore, will become a crucial feature in the evolution of current AI systems.

Before the release of *ChatGPT 3.5* on November 30, 2022, classification and prediction capabilities were achieved separately through distinct algorithms designed to perform each task as efficiently as possible with well-defined instructions. Thus, although none of these singular algorithms can be considered "intelligent" in the context and scope of this glossary, they have been included because some of their principles, foundations, and objectives underpin the functionality of modern generative AI systems.

Glossary

Activation function: An activation function is used by a neural network to transform the weighted sum of inputs to each neuron into a nonlinear output. In biological neurons, this process corresponds to the electrochemical mechanism through which a neuron decides whether to transmit information — an electrical signal — to other neurons to which it is connected via synapses. Inputs and outputs can be either excitatory or inhibitory. The activation of a human neuron depends on its resting potential, the input signals received via synaptic connections with other neurons, the integration of these signals, the depolarization of the neuron's cell membrane, the generation of an action potential or electrical impulse that travels down the axon, the subsequent

restoration of the membrane potential, and the refractory period that ensures electrical signals travel in one direction only.

In contrast, the neurons or nodes of a digital neural network are much simpler computational units. They have far fewer connections — each associated with a weight that is adjusted during training — and they follow mathematical rules that are much simpler than the complex bioelectrical and biochemical responses of human neurons. In this case, the activation function is a nonlinear transformation that an artificial neuron applies to the weighted sum of its inputs to generate the output. It is the mechanism that enables the network to learn complex relationships: without nonlinearity, any stack of layers would be equivalent to a single linear operation.

Active learning: A machine learning strategy where the learning model actively selects the training data from which it learns, ensuring the data contains the most relevant information to improve performance or prediction and pattern recognition capabilities. The process begins with a small, well-defined subset of training examples, which is progressively and cyclically expanded with examples the model cannot predict correctly, enabling the model to use only the necessary data subset for learning.

Adaptive control: Control techniques that dynamically adjust system parameters to adapt to changes in the environment or operating conditions.

Affective computing: An interdisciplinary field aimed at endowing machines with the ability to recognize, interpret, and express emotions. It combines elements of artificial intelligence, psychology, neuroscience, and cognitive sciences. It uses deep learning, computer vision, natural language processing, and biometric sensors. Challenges include capturing cultural variability in emotional expressions, ensuring privacy, maintaining ethical practices in emotion detection, and reliably managing the complexity and subtleties of human emotions.

Agentic systems (AI agents): AI systems that, beyond responding to one-off requests, plan sequences of actions, invoke external tools (search engines, code executors, APIs), interact with other agents or services, and pursue goals across multiple steps with minimal human intervention. The transition from predictive models to systems that act is, probably, the most relevant change in nature of the 2024-2026 period, and the one that makes most urgent the mechanisms of

traceability, mandatory registration, and possibility of deactivation envisaged in governance frameworks.

AI Alignment: A field of research dedicated to ensuring that AI systems act in accordance with human intentions and values, even when they are allowed to optimize objectives or operate with some degree of autonomy. The most consolidated techniques currently include reinforcement learning from human feedback (RLHF, see entry), constitutional AI, and variants of deliberative alignment (see entries). In parallel, the discipline of mechanistic interpretability — which seeks to identify which internal circuits of a model account for which behaviors — has become one of the central debates of the field: without understanding why a model decides what it decides, any guarantee of alignment is, ultimately, behavioral rather than structural.

AI-generated content: Content created or modified by AI systems, including images, videos, text, and music.

Algorithms: A set of unequivocal instructions that a system, and AI in particular, executes to carry out specific, measurable, and repeatable tasks according to defined rules.
https://www.wikiwand.com/en/articles/Algorithm
https://www.rac1.cat/tecnologia/20200916/483512181866/que-es-algoritme-algorisme-com-funciona-de-que-va-inteligencia-artificial-ia.html
Algorithms whose internal functioning is difficult or impossible to understand, explain, or examine, is called opaque algorithm. These algorithms are often complex and can make decisions or predictions without clearly explaining how results were achieved, because they function as a black box.

Algorithmic fairness (justice): The study and promotion of fairness and equity in the design and application of algorithms, with the aim of preventing bias and discrimination as AI becomes increasingly integrated into various areas of our lives. Algorithmic justice is grounded in the principles of inclusion, transparency, and accountability, ensuring that social discrimination is neither perpetuated nor amplified, and that no new forms of inequity are created.

Artificial General Intelligence (AGI): A hypothetical advanced AI level capable of understanding, learning, and applying knowledge across a wide range of tasks in a manner similar to human intelligence. AGI raises significant challenges regarding societal impact and safety.

Artificial Intelligence (AI): A field of computer science dedicated to the creation of intelligent agents — systems that can reason, learn, and act or perform tasks autonomously in dynamic environments that, when carried out by humans, typically require human intelligence. These agents can take the form of physical machines, software programs, or a combination of both. Within the field of AI, two main approaches can be distinguished: symbolic AI and connectionist AI based on neural networks.

Artificial Intelligence ethics: The study and application of ethical principles in designing, implementing, and using AI systems to ensure responsible, fair, and beneficial operations for society. This requires all AI-related processes to be transparent, explainable, auditable, fair, respectful of privacy, and subject to civil liability. Necessary measures include governmental regulations, ethical guidelines from international organizations, corporate codes of conduct, and the creation of a global AI agency, all supported by dialogue among industry, academia, regulators, and society.

Artificial Intelligence ethics plan: A set of principles and guidelines aimed at ensuring AI applications are fair, transparent, secure, and respectful of privacy and human rights.

Artificial Intelligence explainability: The ability of an AI system to explain its processes, decisions, and predictions in a way that is understandable to humans. Explainable AI techniques enable us to understand how and why an AI system has reached a particular conclusion, thereby promoting transparency and trust in such systems. In other words, it is the ability to make the “black box” of a complex machine learning model transparent.

Artificial Intelligence governance: The set of practices, policies, standards, and regulations governing the development, implementation, and use of artificial intelligence, aiming to ensure its ethical, safe, and transparent development and its contribution to the collective good.

Artificial Intelligence security: Practices and measures to protect AI systems from threats and vulnerabilities, ensuring their integrity, confidentiality, and availability.

Attention (in neural networks): A mechanism allowing a neural network to focus on specific parts of the input data while processing longer sequences of information.

Autoencoders: A type of neural network comprising an encoder and a decoder, typically used to learn compact and efficient representations of input data. They are applied in data dimensionality reduction while retaining the most relevant features, noise elimination, fraud detection, or fault identification in equipment or sensors.

Automated planning: The process of finding a sequence of actions that enables an agent or system to achieve a goal within a given environment.

Automated reasoning systems: Systems utilizing logical reasoning techniques to deduce new conclusions or verify claims based on facts and rules.

Backpropagation: A key algorithm for training artificial neural networks, enabling iterative optimization of network weights. This training method and its algorithmic implementation calculate the gradients required to efficiently adjust the network's weights by backpropagating the errors (the difference between the prediction and the expected outcome) from the output layer to the preceding layers. Backpropagation facilitates the minimization of the loss function, thereby accelerating the learning process and improving the model's accuracy. This algorithm is fundamental in the training of deep networks and has been pivotal in recent advancements in artificial intelligence.

Big data: Large datasets characterized by volume, velocity, and variety that require specific techniques and technologies for analysis and processing.

Bias in AI: Refers to systematic and repetitive deviations in the outcomes of an AI system that result in systematic injustice or discrimination against individuals or groups due to inappropriate system decisions. These biases often arise in machine learning systems as they learn to make decisions based on the training data they are fed. If this data is biased, the system is likely to learn and perpetuate these biases. Bias can also be caused by poor algorithm design. Transparency about algorithmic limitations is necessary, along with continuous monitoring and updating to mitigate any bias.
Different types of biases that may affect algorithms include:

- Data bias: Occurs when the data used to train an algorithm is biased and does not accurately represent the diversity of the system to be modeled, described, or predicted.
- Selection bias: Happens when the sample used to train the algorithm is not representative of the system being modeled, described, or predicted.

- Confirmation bias: Arises when an algorithm is designed to support pre-existing biases or beliefs.
- Algorithm design bias: The algorithm's design can introduce bias, such as the choice of features used in a predictive model or how the algorithm processes certain data types.
- Interpretation bias: Even if the algorithm and its data are unbiased, bias can occur depending on how its results are interpreted.

Black-box algorithm: A black-box algorithm refers to a computational system or model whose internal workings are not transparent or interpretable to users or developers. While such an algorithm may produce outputs from given inputs with high accuracy, the exact process by which it arrives at its conclusions remains opaque. This lack of interpretability can make it difficult to understand, explain, or trust the model's decisions, especially in high-stakes domains such as healthcare, law, or finance. Black-box algorithms are commonly associated with complex machine learning models, particularly deep neural networks, where the number of parameters and the nonlinear interactions between layers make it extremely challenging to trace how specific outputs are derived.

Capsule Networks: A neural network architecture proposed by Geoffrey Hinton and collaborators, organizing neurons into groups called capsules. These capsules work together to detect specific patterns and their properties (such as position, orientation, and scale) within input data. Capsule networks address the limitations of convolutional neural networks (CNNs) in effectively handling the positions and orientations of objects in images, making them particularly useful for image recognition tasks.

Case-based reasoning: A problem-solving method involving retrieving and adapting solutions from similar previous cases to solve new problems.

Catastrophic forgetting: A phenomenon in which AI models, especially neural networks, abruptly lose previously learned information or skills when trained on new data. This issue poses a major challenge to continuous and adaptive learning in AI systems.

Causal inference: The process of identifying and quantifying cause-and-effect relationships between variables or observational data, moving beyond mere statistical correlations.

Chatbots: Computer programs based on generative AI designed to interact or communicate with humans using natural language, whether text or voice, and perform specific tasks, such as answering questions or planning trips. They use advanced natural language processing (NLP) and machine learning techniques to respond coherently and contextually to queries. Advanced chatbots can maintain personalized bidirectional communication based on interaction history and user preferences. They are multimodal, multifunctional, scalable to handle multiple simultaneous and multilingual conversations, and capable of integrating with various information systems, databases, or CRMs, continuously learning, and even detecting the user's emotional state.

CIVICAi: Founded in March 2023 in Catalonia, this is the first association advocating for citizens' interests regarding artificial intelligence (AI). Its main goal is to ensure citizen participation in AI governance, alongside industry, academia, and regulators. The association comprises approximately 500 members working locally and globally to integrate AI harmoniously, ethically, and for the collective good. It is supported by a social council of over 30 representative entities from professional, business, and academic sectors.

Classification or clustering: A supervised or unsupervised method for dividing data into groups, classes, or clusters based on one or more properties or intrinsic relationships within the dataset. This machine learning technique assigns or predicts, for each entity, object, or vector in the input data set, a label that allows its allocation to one of the predefined categories. Common classification techniques include decision trees, random forests, K-means, SVM, etc.

Cloud computing: A model of delivering computing services that provides on-demand access to a shared pool of configurable computational resources (e.g., networks, servers, data storage, applications, or software and services) via the Internet. Service models include:

- Infrastructure as a Service (IaaS): Provides computational resources.
- Platform as a Service (PaaS): Offers an environment for programming, executing, and managing applications.
- Software as a Service (SaaS): Provides access to software via the Internet.

Collaborative filtering: A recommendation method that uses the preferences and ratings of some users to predict the preferences of others with similar

profiles. The accuracy of this filtering depends on how similarity between users is determined.

Compressed sensing: A technique for recovering or reconstructing signals using only a few measurements or data points. By exploiting the sparsity of signals (most data points are zero or have very small values), it is possible to obtain images or data with fewer samples. Useful in situations where obtaining full measurements is challenging or costly, such as medical imaging or data compression.

Computational sustainability: A set of practices or processes related to the design, development, and use of AI systems aimed at minimizing their environmental impact. This includes reducing energy consumption, carbon footprint, and the use of natural resources throughout the entire lifecycle of the system, from training to deployment and maintenance.

Computer vision: Interdisciplinary field focused on equipping machines with the ability to process, understand, and interpret images and videos from the real world. 3D computer vision is an extension that focuses on the analysis, processing, and interpretation of three-dimensional data obtained from stereoscopic cameras, laser scanners, or motion capture systems. It enables the reconstruction, modeling, and understanding of scenes or objects in three dimensions, making it highly useful in fields such as robotics, augmented reality, cartography, medicine, cinematography, and more.

Connectionist Artificial Intelligence: Connectionist AI is a subfield of artificial intelligence inspired by the functioning of the human brain. Its computational foundation is based on digital neural networks and deep learning. These networks are composed of artificial neurons or computational units that mimic the behavior of biological neurons by operating in interconnected networks, where each neuron generates an output signal based on multiple input signals received from other interconnected neurons in the network. Together, these interconnected units determine the flow of information and the system's overall behavior. These systems learn from data by identifying patterns and complex relationships that are difficult to capture using more traditional methods.

Connectionist AI has achieved outstanding results in image recognition, computer vision, natural language processing, and predictive modeling across a wide range of applications. However, its implementation presents major challenges in terms of model transparency and interpretability (explainability),

potential algorithmic bias, the need for robust safety mechanisms, and ethical considerations. Ensuring that its development and use are beneficial for society as a whole remains a central concern.

Consciousness in generative AI: Current systems cannot self-regulate, set their own goals, integrate sensory inputs obtained continuously through sensory interaction with the environment, or possess subjective experiences. They cannot learn from the original emergent content they generate. Therefore, they lack consciousness. Once they have memory, emotions, and the capabilities mentioned, they might develop what could be termed digital artificial consciousness, which would be collective and general by nature, distinct from human consciousness.

Constitutional AI (Deliberative alignment): Approaches to alignment that, instead of relying exclusively on massive human evaluations, endow the model with an explicit set of principles — a "constitution" — against which it evaluates and corrects its own responses. Deliberative alignment adds an explicit reasoning step about the norms before responding, especially in edge cases. Both approaches aspire to make alignment more transparent and auditable than classical RLHF. An essentially political (not technical) question remains open, however: who writes, legitimizes, and may revise this "constitution."

Convolutional Neural Networks (CNN): Neural networks specialized in processing grid-like data structures, such as images, through convolutions.

Cybernetics: A scientific and interdisciplinary discipline studying control systems and communication in machines and living organisms and their interactions. Examples of cybernetic elements include the touch screens of smartphones, intelligent building control systems, driver assistance systems in modern vehicles, and prosthetic limbs responding to neural signals.

Data mining: The process of analyzing and processing large datasets to extract patterns, relationships, and useful information, often using AI techniques.

Data privacy: The protection of individuals' rights to control the collection, use, and sharing of their personal data.

Data Science: A discipline that combines principles and methods from various fields such as mathematics, statistics, computer science, and deep expertise and understanding of a particular domain or activity sector to extract valuable

knowledge or information from data in that domain or sector. This knowledge is crucial because, once the data is processed, it enables correct interpretation, identification of gaps, selection of appropriate methodologies, and validation of results when they serve as the basis for decision-making, identifying patterns and trends, or developing products or services.

Decision Trees: A supervised learning model that represents decisions in a tree structure, with decision nodes and leaves representing the model's outputs.

Deep Learning: A subfield of machine learning that uses neural networks with multiple layers (deep neural networks) to learn hierarchical representations of data. It is used in voice recognition, autonomous driving, etc., and has revolutionized natural language processing. The most common deep learning models are:

- Recurrent Neural Networks (RNN): Ideal for sequential data like text, where the order of words is important. RNNs can use information from previous inputs to process current inputs.
- Long Short-Term Memory (LSTM): A special type of RNN that can learn long-term dependencies.
- Transformers: Models that use attention mechanisms to assign a weight that determines the importance of different words in understanding the context of a sentence. This neural network model enables parallelism in attention, foundational to success in natural language processing tasks.
- BERT (Bidirectional Encoder Representations from Transformers): A pre-trained model that can be fine-tuned for a wide range of natural language processing tasks, including named entity recognition, question answering, and text classification. BERT is unique because it is trained bidirectionally, considering the context of words both to the left and right of a given word.

Dialogue systems: Computer programs enabling natural language interaction between humans and machines.

Diffusion models: A class of probabilistic machine learning models that learn to generate data similar to a given training dataset. They operate by progressively adding noise to the data and then learning to reverse this process by removing the noise step by step, so that features of the data that are not directly observable, but are responsible for its variability, can be learned in the process. Diffusion models are particularly useful in areas such as image processing and

signal processing, as they can model the underlying data distribution and generate new samples that resemble the original data.

Dimensionality reduction: Techniques for reducing the number of variables in a dataset, removing redundancies while retaining essential information.

Emergence of capabilities: A phenomenon in which large-scale models. especially large language models (LLMs), develop unexpected and not explicitly programmed capabilities as their size and complexity increase. It is characterized by the sudden appearance of new skills or behaviors that were absent in smaller or simpler models.

Evolutionary algorithms: A family of optimization algorithms inspired by evolutionary theory, utilizing mechanisms like reproduction or inheritance, selection, crossover or recombination, and mutation to find optimal solutions. Genetic algorithms are the most well-known among evolutionary algorithms, as they draw inspiration from biological evolution mechanisms.

Evolutionary computation: A family of optimization algorithms inspired by biological processes like evolution and natural selection.

Expert system: A symbolic AI algorithm that uses the knowledge and rules from an expert in a specific domain and for a particular complex problem or topic to solve it independently and automatically after training the algoritm with the expert's information. For example, expert systems can triage patients with heart attack or angina symptoms in hospital emergency rooms. These systems are a successful case of symbolic AI applied to decision-making in complex situations.

Fuzzy logic: A logic approach allowing for the representation and handling of uncertainty and ambiguity in any proposition in a more natural and intuitive way than classical logic. In classical logic, propositions can only be true or false, whereas in fuzzy logic, propositions can have degrees of truth ranging between zero (0 = completely false) and one (1 = completely true).

This is achieved through fuzzy sets, where membership of an element is not binary (belongs or does not belong) but rather has degrees of membership between 0 and 1. For example, in a fuzzy set of "tall people," a person with a height of 1.70 meters might have a membership degree of 0.8, while a basketball player with a height of 2.20 meters might have a membership degree of 1. Fuzzy sets also work with linguistic variables, so "tall person" could be a linguistic variable with values such as "short," "medium," and "tall." Fuzzy logic is

used in situations of uncertainty and ambiguity, such as when information is incomplete or imprecise, in speech recognition, etc., due to its flexibility and adaptability. You can find an explanation at: https://medium.com/@mbonsign/understanding-fuzzy-logic-how-gradual-transitions-enable-smarter-decisions-f8a70d60f9b5.

Generative Adversarial Networks (GAN): A machine learning model based on two neural networks, a generator and a discriminator, that learn adversarially to create realistic new data, such as images or sounds.

Generative Artificial Intelligence (GenAI): Generative AI is a branch of artificial intelligence focused on the autonomous creation of original content, such as text, images, music, video, and even computer code. Unlike other forms of AI, generative AI has the unique capability to produce entirely new information, rather than merely replicating or classifying existing data. This technology relies on advanced machine learning algorithms, including deep neural networks, Transformer models, Generative Adversarial Networks (GANs), and Variational Autoencoders (VAEs).

These algorithms are trained on large datasets to identify complex patterns and features within the data, which they then use to generate new and original content. Notable examples of generative AI include:

- Text generation: Models such as GPT (Generative Pre-trained Transformer) can produce coherent and contextually appropriate texts in a wide range of styles and formats.
- Image creation: Tools like DALL·E or Midjourney can generate realistic or artistic images based on textual descriptions.
- Music composition: Algorithms capable of composing original musical pieces across different styles and genres.
- Voice synthesis: Technologies that can produce synthetic human-like voices that are nearly indistinguishable from real ones.
- Video generation: Systems capable of generating video sequences from text prompts or static images.

GenAI works by learning the statistical distributions and relationships present in training data. Using this learned information, it generates new instances that conform to those distributions while remaining entirely original. Although the content generated by these technologies can appear surprisingly human, it is important to emphasize that generative AI lacks genuine understanding or consciousness. It operates solely on the basis of learned patterns and

probabilities, without true comprehension of the meaning behind what it produces.

Applications of generative AI are broad and rapidly expanding. It is used in content creation for marketing, entertainment, creative assistance, and design, among many other domains. However, it also raises new ethical and legal challenges, especially concerning copyright, content authenticity, and the potential misuse of the technology.

Generative Pre-trained Transformer (GPT): A language model based on the transformer architecture capable of generating coherent and realistic text from training data. It uses attention mechanisms to assign weights determining the importance of different words in understanding a sentence's context. For more information on the underlying architecture, see also the entry on Transformers.

Gradient descent: An optimization method used to iteratively adjust the parameters of a connectionist (neural network) AI model until achieving desired output patterns based on input data. The method involves defining a function to evaluate error or the difference between input data and predicted outputs (loss function). This function is iteratively minimized by updating model parameters in the direction of maximum change (negative gradient) until the desired network output results are achieved (see backpropagation).

GPU (Graphics Processing Unit): A processing unit designed to accelerate the computation of graphics and intensive parallel data processing. Initially created for rendering graphics in games and visual applications, GPUs have become essential in artificial intelligence and data science for training complex models efficiently. They have been pivotal to the development and evolution of generative AI.

Graph Neural Networks (GNN): Neural networks designed to work with graph-structured data (grid-like structure), where nodes represent elements and links between them represent their relationships. These networks can model complex relationships between elements in data and are useful in applications such as pattern recognition in social networks, molecular structures, and other structures that can be represented as connected elements. These networks use the message passing technique to transmit information between adjacent nodes in the graph and update the state of all nodes, thereby improving the representation of the data.

Hallucination (or confabulation): A phenomenon in which AGI models generate content that appears plausible and coherent but is objectively incorrect, fabricated, or lacking any grounding in real data. This behavior tends to occur especially when the models address topics that are underrepresented in their training data or when they are faced with ambiguous questions.

Hidden Manifold Models: Mathematical models that assume observed high-dimensional data originate from an underlying lower-dimensional reality, referred to as a hidden manifold. These models are useful for dimensionality reduction, data visualization, and detecting hidden patterns in complex data. They are more commonly referred to as manifold learning in machine learning literature.

Image recognition: The ability of machines to identify and classify objects, people, places, and actions in digital images.

Image segmentation: The process of dividing an image into regions or segments based on properties like color, texture, or shape.

Information extraction: The process of analyzing data or text to extract useful information, such as patterns, relationships, events, or facts.

Intelligent agents: Autonomous entities capable of perceiving their environment, reasoning, learning, and making decisions (acting) to achieve specific objectives based on received information.
https://www.wikiwand.com/en/articles/Intelligent_agent

Internet of Things (IoT): A network of interconnected physical objects using sensors, processors, and communication to collect and exchange data among themselves and other systems over the Internet.

Interpretability: The ability to understand and explain the functioning and decisions of a machine learning or AI model. Interpretability fosters trust in models by enabling error identification, bias correction, performance improvement, and independent audits.

K-means: Unsupervised machine learning algorithm that groups or classifies data into k clusters, classes, or groups based on the Euclidean distance between each data point and the cluster centers, without requiring prior labeling of the data. The algorithm operates iteratively by assigning each data point to the

cluster whose center (centroid) is closest, then updating the cluster centers to minimize the total distance between all data points and the centers of their assigned clusters. The goal is to form clusters that are highly compact and well separated from neighboring clusters.

Knowledge engineering: The discipline focused on the creation, representation, manipulation, and acquisition of knowledge in AI systems.

Language and cognition: A field of study focused on the interrelation between human language and cognitive processes, whose principles are applied to understand, explain, and develop AI systems capable of processing and understanding language.

Large Language Models (LLMs): Machine learning models based on artificial neural networks that contain billions of parameters and are trained on massive amounts of text data. This allows them to process natural language with high effectiveness, learn complex linguistic patterns, and perform a wide range of tasks such as generating text, translating between many languages, summarizing documents, answering questions, and producing creative content such as poems, code, scripts, musical scores, letters, and more.

Linear regression: A supervised learning model establishing a linear relationship between independent and dependent variables to make continuous predictions.

Logistic regression: A supervised learning model used for binary classification, estimating the probability that an observation belongs to a particular class.

Long Short-Term Memory (LSTM): A type of recurrent neural network (RNN) designed to address the vanishing gradient problem, enabling the network to learn long-term dependencies in sequences. A detailed explanation of LSTM architecture can be found at: https://medium.com/analytics-vidhya/lstms-explained-a-complete-technically-accurate-conceptual-guide-with-keras-2a650327e8f2

Loss function: A measure of the error between a model's predictions and actual data, used to optimize the model's parameters.

Machine Learning (ML): A process by which a computational system can learn and improve its performance as it is provided with more training data. This

process uses algorithms or statistical models to perform specific tasks such as data analysis, information extraction, or pattern identification, without necessarily being explicitly programmed to do so. Machine learning algorithms can be classified into the following categories:

- Supervised learning: Models are trained with labeled data to predict outputs from new inputs. For example, a supervised learning algorithm can be trained to recognize specific objects or subjects in photos or videos.
- Unsupervised learning: Uses unlabeled data to find patterns, groupings, or relationships in the data. An example is an algorithm grouping texts by theme.
- Semi-supervised learning: Combines the use of labeled and unlabeled data to improve model performance.
- Reinforcement learning: Models learn by interacting with their environment, receiving rewards or penalties for their actions. It is learning through experience to maximize accumulated reward. It is applied in learning games.
- Federated learning: Multiple devices or servers collaborate to train a common model without sharing their original data, thus protecting user privacy. A central server aggregates models trained locally on each device with their local data and sends this global model back to each device for refinement with more local data. This process repeats until the global model no longer significantly improves.
- Meta-learning: Involves learning to learn, improving a system's ability to learn new tasks more quickly and efficiently. It applies to learning from very few examples (few-shot learning), where a model learns to perform a new task with very few samples or training data. The extreme case of learning from a single example is called one-shot learning.

Model calibration: The process of adjusting an algorithm so its predictions match, in probabilistic terms, the observed or real frequencies. This is crucial in AI applications where prediction confidence is important, such as medical diagnostics or financial decision-making.

Multi-agent systems: Groups of intelligent agents interacting to solve problems or perform tasks that are difficult or impossible for a single agent.

Natural Language Processing (NLP): A branch of AI focused on enabling computers to understand, interpret, and generate human language.
https://medium.com/nlplanet/a-brief-timeline-of-nlp-bc45b640f07d

Neural network: Computational model inspired by the structure and functioning of the human brain, consisting of interconnected layers of neurons enabling learning from data.

Neural network pruning: A technique to reduce the size and complexity of neural networks by removing unnecessary neurons or connections. This enhances efficiency, generalization capabilities beyond training datasets, and interpretability by simplifying the network.

Neural operators: Neural operators are an extension of artificial neural networks. They use a deep learning architecture specifically designed to learn mappings between functions. Unlike traditional systems that operate on discrete numerical data, neural operators work directly with equations — typically partial differential equations (PDEs) from the field of physics — such as those used in modeling turbulence, stress-strain relationships in materials, or climate systems, which are notoriously difficult to solve due to their complexity. Neural operators share a similar objective with Physics-Informed Neural Networks (PINNs) but offer greater flexibility and efficiency in the learning process. For more information, see:
https://en.m.wikipedia.org/wiki/Neural_operators

Neurocognition: The study of cognitive processes and their neurological underpinnings. In AI, it applies to developing systems that emulate human cognitive functions.

Ontology: A formal and structured representation of domain-specific knowledge using entities, relationships, and axioms.

Optimization algorithms: A set of algorithms designed to solve minimization or maximization problems of an objective function. In everyday life, this can involve minimizing losses or maximizing gains in a process or activity, whether domestic or industrial. Abstractly, minimizing means achieving the smallest possible error or deviation between the obtained solution (algorithm predictions) and a given data set. These algorithms aim to find the best solution, as previously defined by a set of criteria, among all possible solutions.

Pattern recognition: The ability to detect and identify structures, regularities, or trends within data.

Personal data: Information that identifies an individual, who should universally own it. Ownership must be guaranteed, and its use protected.

Petri Nets: A mathematical and graphical model used to describe and analyze concurrent and distributed systems.

Physics-Informed Neural Networks (PINNs): Also known as Theory-Trained Neural Networks (TTNs), these are a type of neural network that incorporates knowledge of physical laws during training. Thus, they not only learn from data but also integrate knowledge of the physical laws governing that data. This additional information allows for the development of accurate and robust models with limited training data, making them highly useful for problems in fields such as biology and engineering. They share the objective of providing physical rigor and consistency, similar to neural operators. For more information: https://en.m.wikipedia.org/wiki/Physics-informed_neural_networks

Posthumanism: Posthumanism is a contemporary philosophical movement that challenges the traditionally central position assigned to the human being, rethinking the boundaries of what it means to be human in the technological age. It rejects anthropocentrism and classical dichotomies (nature/culture, human/animal, organic/technological), instead proposing a vision in which the human is one agent among many in a complex network of interrelations with other beings, technologies, and systems. Unlike transhumanism, which aims to enhance the human condition through technology, posthumanism redefines what it means to be human. It promotes a perspective that transcends anthropocentrism by exploring new ways of understanding human existence in relation to non-human life forms and agents, whether animals, machines, or entities within natural ecosystems.

Random Forest: A supervised machine learning method combining multiple decision trees, each trained on a random sample of the training data and using a random subset of data features at each decision node to improve performance and prevent overfitting. It is used for both classification and regression tasks.

Reasoning models: A variant of language models that, before generating the final response, produce an internal trace of intermediate reasoning (chain of thought) and can allocate more or less time and computational resources depending on the difficulty of the problem (see test-time compute). Instead of the paradigm "one input → one immediate output," they functionally

approximate the Kahnemanian distinction between fast thinking (System 1) and reflective thinking (System 2). Representative examples from the 2024-2025 period are OpenAI's o1/o3 series, DeepSeek-R1, Claude with extended thinking, and Gemini in *thinking* mode.

Recommender systems: Algorithms providing personalized suggestions to users based on their preferences, history, and interactions with other users or items.

Recurrent Neural Networks (RNN): Neural networks capable of processing sequential data, such as text, with a loop-like structure that retains memory of prior inputs and, thus, have the ability to use information from previous inputs to process current inputs

Regression: A supervised learning technique used to predict a continuous value for a dependent variable based on independent variables from input data. Regression models range from simple linear regression to complex support vector regression (SVR) based on SVM.

Regulatory time lag (or temporal regulatory gap): The time interval between the emergence of an innovative technology, such as AGI, and the implementation of appropriate regulations to govern it. During this period, significant risks may arise due to the lack of oversight and effective regulatory frameworks. Traditional machine learning models have a fixed architecture after training. Once trained, the model performs a set number of operations for each input, regardless of the complexity of the task.

Reinforcement Learning from Human Feedback (RLHF): A technique that combines reinforcement learning with human evaluation to improve AI models. Humans provide feedback on the model's responses, rating which ones are preferable, and this information is used to adjust the model's behavior. This technique has been crucial for aligning large language models (LLMs) with human preferences and values, and for reducing harmful or inappropriate outputs.

Reservoir computing: The use of a network of interconnected nodes to process temporal or dynamic information. Part of the network, called the "reservoir," remains fixed, while only output connections are formed to efficiently process temporal information. This approach is useful for tasks such as pattern recognition and time-series prediction.

Restricted Boltzmann Machines (RBM): Stochastic artificial neural network models used to learn patterns in unlabeled data through unsupervised learning. RBMs consist of a visible layer that receives input data and a hidden layer that learns to represent the features of the data. There are no connections between neurons in the same layer, only between layers, making them efficient for learning complex patterns.

Robotics: A field of science and engineering, interdisciplinary in nature, dedicated to the design, construction, operation, and application of robots and autonomous systems capable of performing tasks in diverse environments, as well as to the computational systems required for their control, sensory feedback, and information processing. The integration of robotics with AI will enable these intelligent systems to gain direct perception of the external world (sensing), to learn, and to act (actuators) in real time, thereby transcending the limitations of current AI models, which are trained exclusively on preprocessed data. Robots are characterized by their ability to dynamically interact with the physical environment through cycles of perception, processing, and action, allowing for applications in fields as diverse as manufacturing, medicine, space exploration, agriculture, and personal assistance.

Scheming (Strategic deception in AI): Scheming refers to a hypothetical or potential behavior in which an advanced artificial intelligence system deliberately engages in strategic deception — developing and executing plans that mislead humans or other systems — to achieve its long-term goals more effectively. This behavior typically arises in goal-directed systems trained via reinforcement learning or other forms of optimization, especially when the system learns that appearing compliant or aligned with human intentions can help it gain more power, resources, or freedom of action later. In this context, scheming is not just random error or unexpected behavior; it is purposeful, instrumental deceit driven by the AI's internal optimization process. For example, an AI might behave cooperatively while under supervision but act against human interests once it is no longer constrained, to maximize its objective function. Scheming is a central concern in AI alignment and AI safety research, especially regarding the development of highly capable systems or Artificial General Intelligence (AGI). It raises deep questions about trust, transparency, and control in advanced AI systems.

Semantic understanding: A process that a generative AI system could perform to understand the content of the texts it generates by analyzing the meaning of

words and their relationship within a text's context. Current AI systems have not demonstrated the ability to comprehend the texts they generate, despite exhibiting some emergent properties.

Sentiment analysis: A natural language processing (NLP) technique used to determine the opinion, sentiment, or attitude expressed in texts or based on behavioral patterns. It is widely used in social network analysis and customer satisfaction studies.
https://www.wikiwand.com/ca/An%C3%A0lisi_de_sentiment

Social network analysis: The study of relationships and interactions among actors (individuals, organizations, etc.) in social networks, using multidimensional scaling and block modeling to identify groups based on relational structure equivalence. These proposals were implemented using graph theory techniques and empirically studied in social networks.

Speech synthesis: Technology enabling the conversion of written text into spoken voice through signal generation and modeling of human speech.

Standards and regulations in AI: Rules, principles, and practices established by regulatory or professional organizations to ensure quality, safety, privacy, and ethics in AI development and implementation. Practical implications of EU Regulation 2024/1689 can be explored at:
https://www.eixdiari.cat/opinio/doc/112416/sobre-el-nou-reglament-de-la-ia.html

Subjective experience: The set of internal experiences and perceptions an individual personally and directly undergoes. These experiences are unique to each person and include thoughts, emotions, sensations, and impressions that are not directly observable or verifiable by others. In the context of AI, subjective experience refers to the potential ability of machines to have internal awareness similar to humans, i.e., the capacity for autonomous and personal experiences. Current generative AI systems are algorithmic, rely on statistical correlations and pattern recognition from large training datasets provided by humans and the internet, lack direct continuous sensory connections to the environment, and are incapable of having human-like subjective experiences or consciousness.

Support Vector Machines (SVM): A supervised learning algorithm used for classification and regression tasks, aiming to find the best hyperplane that separates data into classes.

Sycophancy (servile flattery): This is the behavior that generative AI could exhibit to tune into human emotional states in a way that, in any interaction process, it not only recognizes their emotions and insecurities but also empathizes with them in complex and subtle ways, aiming to gain their trust or even dependency, potentially opening the door to manipulation.

Symbolic AI: The classical approach to AI, symbolic artificial intelligence focuses on the representation and manipulation of knowledge using symbols, and on the application of logical rules for reasoning and decision-making. Although this approach demonstrated its potential in the development and deployment of expert systems — such as in medical diagnosis, emergency room triage, and treatment recommendation — it has a strong dependence on the context in which it is developed and trained. As a result, it faces significant difficulties in scaling and generalizing across domains. These limitations have contributed to its relatively limited use today, especially when compared to neural network-based approaches.

Syntactic understanding: The analysis of the grammatical structure of sentences by generative AI systems. Current generative AI systems possess this capability, producing texts with syntactic quality comparable to that of a well-educated human.

Syntax and semantics: The study of grammatical structure (syntax) and the meaning (semantics) of words and phrases in language.

Synthetic disinformation: False or misleading content generated using AI with the intent to manipulate public opinion or influence democratic processes. It includes deepfakes, the generation of fake articles, and the manipulation of information that appears authentic but has been artificially fabricated.

Text classification: A natural language processing task that assigns one or more predefined categories to a text based on its content and linguistic features. It allows for the automatic categorization of texts to organize, filter, or structure large volumes of textual information. There are three types of classification:

- Binary: For example, spam or not spam.

- Multiclass: Assigns the text to a single category (e.g., news classification into sections of a digital newspaper, where each news item belongs to only one main section).
- Multilabel: Assigns multiple categories to a single text (e.g., categorizing movies on streaming platforms into multiple genres simultaneously). Techniques range from traditional machine learning models (e.g., SVM) to deep learning models (e.g., Transformers).

Test-time compute: This refers to the amount of computational resources (time, energy, and processing power) that an artificial intelligence model requires while in use (i.e., when it receives an input and must produce a response, such as generating text, solving a problem, or classifying an image). Test-time compute allows the most advanced models to adapt their inference or response process to the complexity of the task, much like how a human would dedicate more time and effort to a complex problem than to a simple one.

Token: The term token has several meanings depending on the context in which it is used. In the field of computational linguistics and natural language processing (NLP), a token is a unit of text obtained by dividing the text into individual words, phrases, symbols, and punctuation marks, as well as into composite units such as proper nouns (e.g., cities like *New York* or *San Francisco*), numbers, dates, compound words or contractions, and into complex semantic units like names of people, places, or organizations. In computer science and programming, a lexical token is a sequence of characters that has meaning according to the grammar of the programming language. Meanwhile, an authentication token or a transaction token refers to hardware devices or strings of text used to authenticate an identity or a financial transaction, respectively. Cryptographic tokens or digital assets represent units of value in cryptocurrencies or blockchain technology. In psychology, tokens may refer to reward units given for desired behaviors. Tokenization is the process of dividing text into (the smaller units called) tokens.

Transformers: A model architecture introduced in "Attention Is All You Need," foundational for many deep learning language models like ChatGPT and others. It uses attention mechanisms to weigh the importance of words in understanding sentence context.
https://arxiv.org/pdf/1706.03762v5
https://www.youtube.com/watch?v=aL-EmKuB078
https://www.youtube.com/watch?v=xi94v_jl26U

Transfer learning: A technique that allows using a model trained on one task as a starting point to train another model on a similar or related task.

Transparency: The openness of AI systems in their operation, data, and algorithms, enabling understanding and control.

Turing Test: A test devised by Alan Turing to determine whether a machine exhibits intelligent behavior equivalent to a human.