CIVICAi

CIVIC ARTIFICIAL INTELLIGENCE PROTOCOL

CIVICAi

The use of digital technologies related to mobility and communication has become ubiquitous in our lives across the globe. The COVID-19 pandemic accelerated their adoption and transformed how we work, consume, interact, gather information, and engage with machines and the world. Almost without realizing it, we have found ourselves immersed in the Fourth Industrial Revolution, also known as Industry 4.0, characterized by automation and the flow of information among physical, biological, and digital technologies. The 4.0 technologies with the greatest impact on the physical and biological world include, among others, biotechnology, advanced robotics, 3D printing, new materials, and the Internet of Things (IoT). In the digital realm, the 4.0 revolution encompasses blockchain technology, big data and its analysis, cloud computing, cybersecurity, virtual and augmented reality technologies, and Artificial Intelligence (AI).

Although the aforementioned 4.0 technologies open new possibilities and scenarios for innovation in all areas of society and the economy, they have so far contributed to incrementally, rather than revolutionarily, increasing the efficiency, productivity, and quality of many services and products. The emergence of generative AI (GenAI), which has experienced unprecedented growth in global user adoption and is projected to have a significant impact on productive sectors and public administrations, is likely the catalyst that could enable 4.0 technologies to become a true revolution. This revolution would be characterized by a profound symbiosis between humans and machines. Some experts refer to GenAI as the dawn of Industry 5.0, marked by collaboration between humans and advanced intelligent technologies to solve complex problems and create new forms of interaction and experience.

As with all revolutionary processes, GenAI offers numerous opportunities but also poses new challenges and threats to humanity. We must ensure that humanity's natural tendency to overestimate short-term consequences and risks does not cause us to lose sight of the medium- and long-term risks and impacts of GenAI. The omnipresence of a technology that emulates human abilities carries the risk of humans losing social skills and being manipulated in collective or personal decision-making. This is why CIVICAi advocates for citizen participation in the governance of AI and adopts this protocol for civic artificial intelligence, designed to serve people and promote the common good.

CIVICAi

INDEX

1. CIVIC INTELLIGENCE

One of the most representative technologies of the Fourth Industrial Revolution, and what makes it truly revolutionary, is Artificial Intelligence, with its rapid and continuous development and its cross-cutting impact. The advent of generative AI systems and the impending arrival of Artificial General Intelligence (AGI) represent a revolutionary shift in human evolution and in our understanding of intelligence, language, and cognition. As members of CIVICAi, one of our responsibilities is to facilitate public discourse and enhance social understanding of these profound and rapid changes so that, once they materialize, the new evolution that follows can be assimilated harmoniously and for the collective good. The set of questions and answers proposed in this document is designed to provide a roadmap with sufficient information to navigate the complexities of generative AI, addressing both the technical nuances and some of the broader philosophical implications, albeit in a superficial but sufficient manner.

Historically, our conception of intelligence has been deeply influenced by Cartesian dualism, where René Descartes postulated a strict separation between mind and body[1]. This perspective shaped the early stages of AI, with efforts to emulate human cognition through logical rules and symbolic manipulation, known as symbolic AI. Noam Chomsky's theory of universal grammar further reinforced the idea that the capacity to acquire language is innately programmed in the human brain, emphasizing the primacy of inherent structures over learned patterns[2].

However, recent advancements in AI, particularly through the work of researchers like Geoffrey Hinton, Yoshua Bengio, and Yann LeCun, recipients of the 2018 Turing Award, have challenged these traditional and orthodox linguistic postulates[3]. The pioneering work of these researchers in neural networks and deep learning has demonstrated that large-scale language models (LLMs) can achieve remarkable proficiency in understanding grammar or syntax in texts and generating human-like language. This connectionist perspective aligns with the philosophical work on mind and language by Ludwig Wittgenstein, which emphasizes the use of language in specific situations and contexts rather than exclusive reliance on fixed grammatical structures and rules. This implies that

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

understanding language requires attending to the social contexts in which it unfolds, rather than to abstract, invariant linguistic structures[4].

The heterodox perspective advocated by Geoffrey Hinton, Yoshua Bengio, and others suggests that learning from large amounts of data, rather than relying on pre-programmed rules, can lead to a form of practical understanding that challenges Chomsky's postulate of innate structures. Recent developments in neurocognitive mechanisms, based on a mechanistic view of the mind and advanced by Gualtiero Piccinini, further support this argument, indicating that the mechanisms underpinning human cognition can be represented analogously, yet differently, in artificial systems[5]. It is important to note, however, that the processes through which current generative AI produces original semantic content are algorithmic. They rely on statistical correlations and pattern recognition derived from large training datasets provided by humans and the Internet. Therefore, these processes differ fundamentally from the semantic processes of the brain, which are inherently biological, context-dependent, intentional, self-regulating, and integrate sensory inputs, memory, emotions, and other cognitive functions tied to the continuous interactions between the mind-body unit and its environment—characteristics associated with what we call consciousness.

As we move closer to a new era shaped by artificial intelligence, with its potential not only to imitate but also to enhance human cognitive abilities in unprecedented ways, it is crucial to recognize that every revolution involves breaking away from the prevailing orthodoxy. The success of connectionist AI (neural networks) represents not only a technological advance but also a paradigm shift in how we conceptualize intelligence itself. By adopting a heterodox perspective that integrates ideas from multiple disciplines, we will be better equipped to address the ethical, social, and philosophical challenges posed by the advanced capabilities of generative AI and future AGI.

A fundamental aspect in this context is the computability of intelligence. Historically considered an exclusively human characteristic—albeit evolutionary—intelligence is now perceived as an emergent property that could also arise in complex digital systems, such as neural network algorithms. These algorithms powering generative AI lead us to question the limits of the computability of intelligence and whether it can be fully reproduced or emulated by machines.

---

[4] https://philosophynow.org/issues/106/Wittgenstein_Frege_and_The_Context_Principle
[5] https://www.thebsps.org/reviewofbooks/gualtiero-piccinini-physical-computation/

This opens the door to discussions about the drawbacks and risks associated with generative AI. In the short term, these risks may include bias, violations of privacy and intellectual property, ethical concerns, rapid labor market transitions, orchestrated disinformation, the erosion of democratic values, or even the alteration of democracy itself. AI systems can reinforce existing biases if training datasets are not adequately supervised. Moreover, the mass collection of data raises privacy concerns, and the use of AI-generated content presents challenges regarding intellectual property and copyright. It is crucial that large language model (LLM) providers operate within a legal framework—ideally as universal as possible—that compels them to mitigate negligent discourse and align their models with verifiable or "true" facts through open and democratic processes. This would help ensure that these technologies serve society responsibly and transparently, respecting fundamental rights and promoting collective well-being[6]. The long-term 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, generative AI—surpasses human intelligence. This prospect raises profound philosophical, ethical, and existential questions about humanity's role and relevance in a world where machines might exceed human cognitive abilities[7]. This would imply that generative AI or AGI could autonomously manage its value function or criteria for achieving objectives, which might not align with human interests or goals. Long-term risks also raise the idea of a post-humanity, where the integration of advanced AI into human society transforms human experience and identity without the full capacity to have democratically decided upon such changes.

It is essential to ask: who will oversee, and how will current and developing generative AI systems, which are predominantly in the hands of private providers, be effectively supervised? The challenge is daunting, as existing legislation and regulations are not global and are limited to establishing a punitive framework that acts retroactively to halt or correct malicious actions only after they have occurred and spread. Therefore, beyond deterring providers with sanctions, states should globally agree on a real-time monitoring framework encompassing data with multimodal inputs, training processes, algorithms, and the outputs of generative AI systems and future AGI. This framework should be implemented through a network of state-owned

[6] https://doi.org/10.1098/rsos.240197
[7] https://lab.cccb.org/en/the-singularity/

computational centers equipped with hardware and software resources equal to or exceeding those held by private companies.

The efficiency and sustainability of generative AI are other critical aspects to consider. Current models require a significant amount of computational and energy resources, which may not be sustainable in the long term, either due to a lack of energy resources or because of conflicts with other human priorities. Hybrid computational solutions, combining analog and digital systems, or chips integrating different types of cores or specialized processors, with optimal integration between hardware and software—similar to the processors in some smartphones—could offer a way to mitigate these issues by improving computational efficiency and reducing the ecological footprint.

In conclusion, this protocol or conceptual and operational framework aims to inspire the public contributions of CIVICAi associates and improve the general public's understanding of AI. Although science is about asking questions rather than providing definitive answers, we dare to offer explanations in the form of measured but well-founded responses to some of the questions humans ask about AI. The goal is to promote, through CIVICAi, well-informed public participation in the governance of AI. We seek to contribute to building an informed, reflective, and respectful discourse that helps society work toward a future where artificial and human intelligences coexist and complement each other in transformative ways for the collective good. This vision is plausible because, when artificial digital consciousness emerges, it will, by its nature, be collective and general.

2. FREQUENTLY ASKED QUESTIONS AND POSSIBLE ANSWERS

ON AI'S UNDERSTANDING CAPABILITIES

1. Does a large language model (LLM) or generative AI, such as ChatGPT 4, Claude, or Gemini, understand and comprehend what it responds to when asked about a specific topic or requested to comment on a subject from any discipline?

Answer: The syntactic quality and the exceptionally high level of processing of syntactic patterns in the responses provided by LLMs, comparable to the language of educated humans, indicate that they possess syntactic understanding. Discussions and disagreements among experts become more pronounced when the question shifts to semantic understanding—whether LLMs truly understand and comprehend the content of their responses—given that they lack subjective experiences involving sensory perceptions, which might be connected to emotions.

2. Do LLMs like ChatGPT 4o comprehend the content of their responses or possess semantic understanding?

Answer: Some commentators or pundits on science and technology, researchers in the field of symbolic AI, and traditional linguists argue that the responses generated by LLMs are merely based on statistical correlations, that LLMs lack the innate linguistic structures of humans, and some even label generative AI as a "stochastic parrot."

However, other scientists and researchers in the field of connectionist AI (neural networks) and experts involved in recent studies in cognitive sciences argue that LLMs exhibit highly complex emergent behaviors, possess generalization capabilities, and that the architecture of neural networks can emulate the role of the neocortex and subcortical structures of the human brain. They conclude that if LLMs exhibit functional behaviors equivalent to those of humans, they must possess some form of understanding, even if only at a practical and functional level, limited by their inability to perceive their environment in real time or to have subjective experiences and, therefore, emotional and cognitive processes.

The theory of **Neurocognitive Mechanisms**, proposed by Gualtiero Piccinini, is a framework in the field of philosophy of mind and cognitive sciences that seeks to explain human cognition through the physical mechanisms of the brain. Piccinini argues that cognition or cognitive processes—such as thinking, memory, and perception—can be understood and explained as a set of

computational processes implemented by neural mechanisms in the brain. These processes are not merely mathematical abstractions but have a physical basis. Neurons and neural networks perform physical computations that result in cognitive phenomena. This mechanistic approach directly challenges Descartes' dualist perspective, which separates the mind (*res cogitans*) from the body (*res extensa*), and indirectly confronts the alignment of symbolic AI with Cartesian dualism.

3. What is syntactic understanding in LLMs, and how does it differ from semantic understanding?

Answer: Syntactic understanding refers to the ability to process and generate language following correct grammatical and structural rules. Semantic understanding involves grasping the meaning and content of the language. LLMs can demonstrate advanced syntactic understanding, but their capacity for semantic understanding is more contested. These models are based on statistical correlations, although they exhibit emergent complex behaviors when trained on a large scale. While LLMs lack semantic understanding in the deep and conscious sense that, according to the theory of Neurocognitive Mechanisms, could be attributed to human neural mechanisms, these models can simulate certain aspects of semantic understanding thanks to their processing and text-generation capabilities.

ON CREATIVITY

4. Can LLMs generate original ideas, or do they merely repeat what they have learned?

Answer: LLMs can combine information in new and unexpected ways, creating content that, if produced by a human, would be considered original and not plagiarism in the strict sense of the term, as it results from learning based on external data rather than a direct copy of it. Everything they generate is rooted in the information and patterns, both evident and subtle, from the large volumes of data on which they were trained. Their "originality" or emergent behavior is the outcome of advanced combination and permutation of existing data, much like a significant portion of the contributions or proposals made by humans, despite the limited processing and memory capacity of the human mind-body unit.

In fact, ChatGPT has passed the Turing Test[8], with its responses being indistinguishable from those of a human when interacting with a human judge

[8] https://www.nature.com/articles/d41586-023-02361-7

unaware of which is which. However, this applies only when the questions are not overly complex or formulated in a highly long-term context, and provided that the human judge is not an expert on the topic and the subject matter does not require a large-scale, highly interdisciplinary context.

5. Can LLMs be creative?

Answer: LLMs can generate creative content such as poetry, art, and music by combining elements in new and interesting ways. However, their creativity differs from human creativity, as it is not driven by personal experiences, emotions, or conscious intentions. LLMs process large volumes of data, which raises concerns about privacy, copyright, and data protection. It is important to ensure that the data used to train these models is collected ethically and safeguarded against unauthorized access or misuse of authorship.

ON AI'S LIMITATIONS

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

Answer: The cognitive limitations of current generative AI models include, among others, the lack of deep semantic understanding of the concepts they process, as their responses are based on statistical patterns in training data. They depend heavily on the quality, quantity, and diversity of the training data, which makes them susceptible to generating incorrect or biased information. They also face challenges in generalizing knowledge, as they are not yet transversal (multifunctional and multimodal; AGI). Additionally, they have limited capacity to resolve ambiguities, demonstrate "common sense," perform logical reasoning, establish deep causal relationships, and they lack consciousness and subjective experience.

On an operational level, the most significant and immediate limitations and risks that can impact the implementation, efficiency, and improvement processes of generative AI models are related to their high and growing computational and energy resource requirements. These arise from the need for scaling, periodic retraining with new data, and algorithmic improvements. Security and privacy systems are necessary to protect them from unwanted attacks, and strict, adaptive control of biases and assurance of fairness are essential.

Decisive steps must be taken to develop and implement hybrid computational solutions (analog-digital), or architectures using chips with specialized processors, with optimal integration between hardware and software, to achieve significantly higher energy efficiency.

ON EMOTIONS AND SUBJECTIVE EXPERIENCES

7. What is a subjective experience?

Answer: A subjective experience is the intact and meaningful understanding derived from an individual's experience, encompassing both the emotional and cognitive impact that directly affects them. It involves how a person perceives and interprets an event or series of events they have witnessed or processed in some other way. This understanding includes the emotions generated and the cognitive reflection on what has occurred, thereby forming a personal and unique interpretation of the reality experienced.

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 because they lack subjective experience. However, they can simulate intelligent behaviors and assist in solving complex problems by analyzing large volumes of data, identifying patterns and trends, generating possible solutions based on historical data, and facilitating collaboration by synthesizing information from various sources. It is possible that a "digital artificial consciousness" could emerge when AI systems are equipped with sensors, learn and interact in real time with their environment and various contexts, and also learn from the content they themselves generate. Such artificial consciousness would likely be collective due to its digital nature.

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 artificial intelligence"?

Answer: Artificial General Intelligence is a theoretical concept, as no AI system currently exists 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 human-like cognitive abilities, including understanding and consciousness. Weak artificial intelligence, on the other hand, refers to systems designed to solve specific problems or perform concrete tasks without any form of consciousness or general understanding.

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, most LLMs do not learn in real time from their interactions with humans. Learning is typically done offline, using large amounts of pre-collected data. However, ongoing research is focused on developing models that can adapt and continuously learn from real-time interactions with humans.

ON ETHICAL IMPLICATIONS

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 negligent outputs and align their models with verifiable facts through open and democratic processes[9].

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 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 ethical implications of using LLMs in scientific research

Answer: The use of LLMs in scientific research can accelerate processes such as literature review, hypothesis generation, and even the planning, execution, and evaluation of tasks and new experiments, with minimal human intervention. For this reason, they carry significant ethical implications by posing risks such as the generation of false but credible citations or data, which could compromise the integrity of research and the consistency of its practical applications. Additionally, the use of these models could amplify existing biases in scientific literature if information is not carefully managed, perpetuating prejudices and inequalities.

Questions also arise regarding authorship and the recognition of LLMs' contributions 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 using these models, including transparency in their application and rigorous verification of generated results to prevent the dissemination of incorrect or misleading data. This will become even more necessary as the prescriptive capabilities of generative AI are deployed and autonomous laboratories come into operation.

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

ON AI BIASES AND HOW TO ADDRESS THEM

14. What are the biases in AI models, and how do they arise?

Answer: Biases in AI models refer to systematic tendencies or prejudices in the model's predictions or decisions, like how we refer to conscious or unconscious human biases related to gender, class, or race. They arise from imbalanced training data, design decisions in the model, and human factors involved in data collection and labeling. Just as we advocate for equal and inclusive education, we must also demand that LLMs be trained with ethical values and in an inclusive manner.

15. How can biases in LLMs be mitigated?

Answer: Mitigating biases in LLMs requires a combination of strategies, including curating diverse and balanced training data, fine-tuning models to identify and address biases during development or training, and implementing post-deployment monitoring and regulation mechanisms to detect and correct issues in the algorithms. This process aims to improve the quality and selectivity of training data. The latter strategy is particularly necessary but challenging, as it demands computational resources comparable to those supporting the LLMs themselves.

16. What are the risks associated with LLMs in terms of misinformation?

Answer: LLMs can “confabulate” (“hallucinate”) and generate false or misleading content convincingly, which can amplify misinformation. These risks can be mitigated through fact-checking mechanisms, algorithmic transparency, traceability of data sources, and collaboration with experts in data verification.

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

Answer: Verifying and validating the results generated by LLMs presents several significant challenges. Firstly, the probabilistic nature of these models means they can produce responses that seem plausible but are incorrect. Additionally, the complexity of the models makes it difficult to understand how a specific response or text is generated, complicating the traceability and explanation of the processes that lead to the model's output.

There is also the phenomenon known as "hallucination" or "confabulation," where models generate information that appears coherent but is not based on verifiable or factual data. Real-time verification of texts and sources for large volumes of generated content poses a significant challenge due to the high computational resources required and the reliance on large data centers, most

of which are privately owned by the companies commercializing generative AI models.

Addressing these challenges requires advanced automatic verification tools, robust fact-checking systems, and the integration of human expert knowledge in the validation process. It is also crucial to develop transparent methodologies that allow auditing and understanding the inner workings of LLMs.

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

Answer: The main challenges in regulating generative AI include:

- Technological advancements, particularly in LLMs, evolve at a pace far exceeding lawmakers' ability to effectively regulate them and to continuously and effectively adapt relevant legislation. Moreover, as systems become autonomous, they will be more capable of bypassing human control [10].
- The global nature of the Internet, which complicates the enforcement of national regulations and makes global regulation essential. Such regulation must involve governments, experts, tech companies, and society at large 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 generative AI 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.
- The need for specific and ongoing training of regulators in generative AI 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.

ON EQUITY AND DEMOCRATIC GOVERNANCE

19. How can equitable access to generative AI 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

[10] https://civicai.cat/wp-content/uploads/2024/05/Managing-extreme-AI-risks-amid-rapid-progress.pdf

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 generative AI 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, develop, and implement 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 LLMs 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.

In the public sector, generative AI can assist in policy development, citizen engagement management, designing and implementing actions in response to public inquiries, and improving the quality and diversity of public services by

analyzing social and economic data. In the private sector, LLMs can be used for market analysis, strategic decision-making, and enhancing the operational efficiency of organizations. However, the incorporation of AI into these processes raises concerns about transparency and accountability, particularly when decisions based on these systems have a significant impact 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 to implement mechanisms for human supervision and set clear ethical frameworks for the use of LLMs in organizational decision-making. This aligns with the regulations set forth in the EU Artificial Intelligence Act, published on July 12, 2024[11].

ON EDUCATION, ART, LANGUAGE, AND CULTURE

22. How can the use of LLMs affect education?

Resposta: The impact of LLMs on education will be significant and rapid, not only because of the extensive use already made by most students, from secondary education to higher education, but also because teachers will need to adapt their tools and learning resources to promote more constructivist learning processes [12]. It should be noted that LLMs can provide personalized assistance to students, adapt to their individual needs, generate resources and educational materials tailored to each learning pattern, and facilitate access to original publications written in different languages, either directly or through artificially generated summaries.

The use of these personalized generative AI assistants presents significant challenges, such as the potential overreliance on these tools, which could affect the development of essential human skills, such as critical thinking, teamwork, problem-solving, and innovation. Regarding teachers[13], the use of LLMs can lead to lesson plans that fail to effectively build students' knowledge, tutoring sessions that may confuse students with incorrect answers, and educational materials based on erroneous concepts. In light of this, it is essential for educators and educational institutions to develop policies ensuring that AI-generated tools are rigorously evaluated and verified and are integrated ethically and effectively into the educational system. This ensures a balance between the use of technology and the need to develop human skills within a

[11] https://artificialintelligenceact.eu/the-act/
[12] https://www.wikiwand.com/ca/Constructivisme_(pedagogia)
[13] https://www.cognitiveresonance.net/resources.html

framework of strict respect for fundamental rights[14].

Nevertheless, neither the lack of clear policies nor the challenges posed have prevented higher education institutions from developing and positively evaluating classroom activities specifically designed to enhance critical thinking. These activities focus primarily on the process of asking incisive and profound questions, evaluating information to draw logical conclusions, and understanding complex topics[15]. These experiences, along with others carried out by members of CIVICAi to foster critical thinking in universities, suggest that the use of LLMs in classrooms could be framed within a methodology rooted in maieutics[16], employing a teaching format akin to that of the ancient Socratic school, under the guidance of each professor. This open and participatory format would facilitate reflection and critical thinking, promoting in-depth discussions and the exchange of ideas between students and teachers.

With students having intelligent personal assistants in their pockets, this shift in approach could enrich the educational experience, foster a more collaborative and student-centered education, and encourage more personalized and dynamic assessment systems. This approach would not only help mitigate the risks associated with overreliance on AI technologies but also promote an educational context where critical reflection and intellectual debate are central. It would ensure that students develop the necessary skills to verify, interpret, and responsibly and ethically use complex information. At the same time, it could make the vertical silos of individual disciplines more permeable, evolve the medieval structure of universities, and return knowledge to where it originated: the process of asking questions to construct understanding.

23. Can LLMs understand complex cultural and social contexts?

Answer: Current LLMs can recognize and generate language within cultural and social contexts based on the data they have been trained on, but their understanding is superficial and rooted in statistical patterns. This can lead to errors in situations requiring a deep understanding of cultural or social contexts. The limitations of generative AI discussed in question #6 are also relevant to answering this question #23.

[14] https://rm.coe.int/artificial-intelligence-and-education-a-critical-view-through-the-lens/1680a886bd
[15] https://civicai.cat/wp-content/uploads/2024/05/Leveraging-chatgpt-for-enhancing-critical-thinking-skills.pdf
[16] https://ca.wikipedia.org/wiki/Mai%C3%A8utica?wprov=sfti1#

24. How can LLMs impact linguistic and cultural diversity?

Answer: LLMs can impact linguistic and cultural diversity in several ways. On the one hand, they can serve as a powerful tool for preserving minority languages by generating content and offering machine translation, helping to revitalize endangered languages and keep cultural traditions alive. On the other hand, there is a risk that they may reinforce the dominant role of majority languages, such as English, as most models are trained primarily on data in these languages, reducing the visibility and use of minority languages. Moreover, LLMs might influence how ideas are expressed in different cultures, potentially homogenizing diverse cultural expressions and eliminating important nuances. To mitigate these risks, it is essential that the development of these models includes diverse cultural and linguistic data and involves close collaboration with cultural stakeholders from the affected languages to ensure a harmonious and respectful approach to all cultures.

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

Answer: LLMs can be valuable tools for preserving and studying intangible cultural heritage. They can assist in processing and analyzing large volumes of cultural data, including oral histories, traditional songs, and cultural practices. They can help transcribe and translate endangered languages to facilitate their preservation and study. Additionally, they can generate interactive representations of cultural practices to promote greater awareness and appreciation of cultural heritage. However, it is crucial to involve cultural communities in this process to ensure accuracy and respect for traditions. Issues of intellectual property and consent must also be addressed to ensure that the beneficiary communities retain control over how their cultural traditions are collected, used, and disseminated.

26. How do or might LLMs affect artistic creativity and cultural production, and what ethical, legal, and socioeconomic implications could arise in the short and long term?

Answer: Generative AI has raised concerns across most 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 ability to influence cultural production across the board can help reduce technical barriers and diversify creative resources. LLMs' capability to introduce interactive and personalized forms of art can not only transform the artistic

experience but also reshape perceptions of authenticity and the value of artistic works.

However, these transformations also bring significant challenges. Ethically and legally, complex questions arise about originality, authorship, and the intellectual property rights of AI-generated works. The labor market in the arts sector could face profound restructuring due to the potential displacement of certain creative roles and the emergence of new hybrid professions combining human and AI collaboration. In this context, fostering collaboration between human artists and AI will be crucial to ensure that AI complements rather than replaces human creativity. There is also a risk of homogenization in artistic production, as well as shifts in the economic valuation of art and creativity.

To address these challenges, it will be essential not only to develop ethical and legal frameworks regulating these new dynamics but also to conduct long-term research and assessment of the impact LLMs will have on cultural diversity and artistic expression. Encouraging balanced collaboration between humans and AI, and educating the public about the capabilities and limitations of AI-generated art, will be fundamental to ensuring a future in which technology enriches rather than limits cultural expression. Finally, protecting the rights of artists will require studying potential consequences, implications, or even compensation in this new creative environment. This revolution compels us to ask fundamental questions about the nature of creativity, the preservation of cultural heritage, the evolution of cultural identities, and the future we envision for human cultural expression in the era of generative AI, which is only just beginning.

ON SUSTAINABILITY AND HEALTH

27. What are the implications of large language models (LLMs) in climate change?

Answer: Large language models (LLMs) have a significant environmental impact due to their high energy consumption, particularly during training and operation phases. To mitigate this impact, it is crucial to adopt strategies that reduce the energy consumption associated with these models. Such strategies include developing more computationally efficient models, using renewable energy to power data centers, optimizing algorithms to minimize required computational resources, and utilizing specialized hardware such as AI-adapted chips. Additionally, exploring emerging technologies like hybrid or analog computational systems could offer more energy-efficient solutions. It is worth

noting that the energy consumed by current generative AI systems exceeds the energy consumption of some of the 193 UN member states.

However, LLMs can also be valuable in combating climate change. They are capable of analyzing large volumes of climatic and environmental data to identify patterns and trends, making predictions about extreme weather events (such as hurricane trajectories) quickly and effectively[17,18], and improving the accuracy of existing climate models. This could help better understand the effects of greenhouse gas emissions and other anthropogenic factors. Moreover, LLMs can be used to evaluate the impact of different environmental policies and offer data-driven recommendations for more effective climate change management. They can also help in effectively communicating climate science to the general public, thereby contributing to the adoption of effective environmental policies.

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

Answer: LLMs can be a powerful tool for detecting and preventing public health crises. Their 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. LLMs should also be rigorously integrated into public health systems, particularly epidemiology services, with clear protocols for verifying and disseminating AI-generated information.

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

Answer: Large language models (LLMs) can profoundly transform healthcare systems, enhancing both patient experience and the detection and treatment of diseases across primary, specialized, and hospital care. Regarding patient

[17] https://www.wired.com/story/ai-hurricane-predictions-are-storming-the-world-of-weather-forecasting/
[18] https://www.freethink.com/robots-ai/ai-based-weather-forecasting

experience, a key aspect is the quality of interaction and empathy shown by healthcare professionals during consultations. LLMs can help reduce the burden on professionals by automating routine tasks, allowing them to focus more on human interaction. For instance, generative AI can assist in the automatic documentation of patient information, transcribing consultation reasons or symptoms described by patients in their own words. This documentation can be directly integrated into their medical records, subject to review by healthcare professionals, with AI suggesting appropriate actions such as referrals, hospital admissions, or treatments. This automation would not only improve efficiency but also enable healthcare professionals to dedicate more time to empathetic patient care, thereby elevating the overall quality of medical attention.

In terms of disease detection and treatment, LLMs can analyze large volumes of multimodal data, such as medical images, electronic health records, and sensor data, to identify patterns that may go unnoticed by humans. This capability is particularly valuable in critical settings like Intensive Care Units (ICUs), where real-time analysis of diverse data sources can generate alerts and pre-alerts before significant health deterioration occurs, facilitating early intervention. These capabilities can significantly improve risk management and reduce preventable adverse events.

Additionally, LLMs can optimize workflow and resource management, including human, financial, and equipment resources, particularly in nursing services. By analyzing historical and real-time health system data, LLMs can enhance workforce planning, considering workloads, skills, preferences, and individual profiles of nursing professionals. This would reduce human errors and identify opportunities for improvement. Automating repetitive and administrative tasks, such as data entry, appointment scheduling, and medication follow-ups, as well as improving team coordination, would facilitate more efficient management while increasing safety and quality of patient care.

Finally, adopting generative AI in healthcare systems should involve collaboration between health systems across different countries. This would enable the secure sharing of anonymized data, diagnoses, treatments, and clinical outcomes, accelerating global medical advancements and improving responses to public health crises on a global scale.

In summary, integrating LLMs into healthcare systems has the potential to significantly improve both patient care and clinical efficiency. However, it is crucial to ensure ethical and secure use of these technologies, protecting medical data privacy and ensuring that automated decisions are based on AI

agents subjected to randomized controlled trials, with the participation and supervision of medical professionals to assure reliability.

ON WORK: CHALLENGES AND OPPORTUNITIES

30. How can generative AI transform workplaces?

Answer: Generative AI can automate tasks that require language processing, such as text writing, document analysis, and creative content generation, affecting all industries and most professions. However, it will also create new job opportunities in areas such as AI development, data management, cybersecurity, and AI system oversight, among others.

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 LLMs for content creation and management on social media platforms?

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

Transparent regulatory policies on the use of AI-generated content, robust mechanisms to detect deepfakes and misinformation, and user education about the presence and limitations of AI-generated content on these platforms are essential. Promoting collaboration between social media platforms, regulators, and civil society is also crucial to effectively addressing these challenges.

33 What are the biggest technical challenges in developing LLMs?

Answer: The technical challenges that generative AI developers must overcome to advance these models toward more general intelligence can be identified and classified based on their likelihood of realization in the short (1-2 years), medium (more than 2 years), and long term (more than 4 years). These classifications are tentative, as the predictability of complex, non-linear, and rapidly evolving systems is low.

- Short term (1–2 years): Improvements in algorithms and hardware architectures to reduce the time and resources needed to train and run LLMs; reduction of energy consumption and the corresponding carbon footprint; enhancement of LLM interpretability; better management of training data; and adaptability to specific domains or areas of knowledge without losing general information.
- Medium term (more than 2 years): Advanced multimodality to effectively integrate different input and output modalities (text, image, audio, and video) into a single generative AI model; continuous learning without the need for complete retraining; improved ability to perform complex and abstract reasoning beyond simple statistical association; incorporation of advanced security systems to protect data privacy and prevent malicious use; and personalization without compromising efficiency.
- Long term (more than 4 years): Full contextual artificial awareness that enables generative AI to achieve deep and dynamic understanding of cultural, temporal, and situational contexts; autonomous real-time learning without human intervention; causal reasoning to understand and model 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) with all its potential capabilities, including integration of functionalities, cognitive flexibility, deep contextual understanding, metacognition, levels of self-awareness, and more.

ANNEX. 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.

Activation function: A function used by neurons in a neural network to transform the weighted sum of inputs into a nonlinear output. In human neurons, this corresponds to the biological electrochemical process by which a neuron decides what information or electrical signal to transmit to other neurons connected via synapses. In artificial neural networks, these activation functions follow simpler mathematical rules and have limited interconnections compared to biological neurons.

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.

Algorithms: A set of unequivocal instructions that systems in general, and AI in particular, use to perform specific, measurable, and repetitive tasks according to a set of rules or instructions. Given initial conditions, an algorithm carries out a sequence of pre-established instructions to achieve an objective characterized by a set of final conditions.
https://www.wikiwand.com/ca/Algorisme
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: The study and promotion of equity in algorithm design and application to prevent biases and discrimination as AI is increasingly integrated into daily life. Algorithmic fairness emphasizes inclusion, transparency, and accountability.

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 creating intelligent agents—systems capable of reasoning, learning, and acting autonomously in dynamic environments. These agents can be physical machines, software, or a combination of both. AI includes symbolic approaches and connectionist approaches 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 to understand and explain the results and decision-making processes of an AI model in a way comprehensible to humans. In other words, it is the ability to make the "black box" of a complex machine learning model transparent.

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

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.

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.

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.

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.

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.

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 university 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.

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 AI: A subfield of AI inspired by the functioning of the human brain, relying on digital neural networks and deep learning. These systems learn from data by identifying patterns and complex relationships. Connectionist AI excels in tasks like image recognition, natural language processing, and predictive analysis but faces challenges regarding transparency, algorithmic bias, safety, and ethical deployment.

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.

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 designed to learn and generate data similar to a training dataset. They operate by adding noise to data and then gradually removing it, enabling the learning of unobservable data features responsible for variability. Useful in areas like image processing and signal analysis.

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

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/@javierdiazarca/lógica-difusa-ejercicios-propuestos-b99603ef1bc0.

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

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.

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.

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.

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).

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/ca/Agent_intel%C2%B7ligent

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: An unsupervised learning algorithm that groups or classifies data into k clusters based on the Euclidean distance of each data point to the cluster centers. The algorithm iteratively assigns data to the nearest cluster center and updates cluster centers to minimize the total distance between data points and their respective cluster centers.

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 with billions of parameters, trained on extensive text datasets. LLMs excel in natural language processing tasks, such as text generation, translation, summarization, question answering, and creative writing (e.g., poems, codes, scripts).

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 networks: Computational models 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: An extension of artificial neural networks designed to learn and transform functions in specific ways. Unlike traditional systems that handle numerical data, neural operators work with equations, often in partial derivatives, to model phenomena like turbulence, stress-strain in materials, or

climate studies. They share goals with Physics-Informed Neural Networks (PINNs) but provide additional flexibility.
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

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.

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 focused on the design, construction, operation, and application of robots and autonomous systems capable of performing tasks in diverse environments.

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.

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

Servile flattery (sycophancy): 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 emphasizes them in complex and subtle ways, aiming to gain their trust or even dependency, potentially opening the door to manipulation.

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.

Symbolic AI: A classical AI approach focusing on representing and manipulating knowledge through symbols and applying logical rules for reasoning and decision-making. While successful in applications like expert systems for medical diagnosis, it struggles with scalability and generalization, limiting its current use compared to neural networks.

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.

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).

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.