Artificial Intelligence in Science and Society: the Vision of USERN

Tommaso Dorigo, Gary D Brown, Carlo Casonato, Artemi Cerdà, Joseph Ciarrochi, Mauro Da Lio, Nicole D'souza, Nicolas R Gauger, Steven C Hayes, Stefan G Hofmann, et al.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Artificial Intelligence in Science and Society: the Vision of USERN

TOMMASO DORIGO[1,2,3], (Ed.), GARY D. BROWN[4,3], CARLO CASONATO[5], ARTEMI CERDÀ[6,3], JOSEPH CIARROCHI[7], MAURO DA LIO[8], (Member, IEEE), NICOLE D'SOUZA[9,10], NICOLAS R. GAUGER[11], STEVEN C. HAYES[12,13,3], STEFAN G. HOFMANN[14], ROBERT JOHANSSON[15], MARCUS LIWICKI[1], FABIEN LOTTE[16,3], JUAN J. NIETO[17,3], GIULIA OLIVATO[5], PETER PARNES[18], GEORGE PERRY[19,3], ALICE PLEBE[20], IDUPULAPATI M. RAO[21,22,3], NIMA REZAEI[23,3], FREDRIK SANDIN[1], ANDREY USTYUZHANIN[24,25], GIORGIO VALLORTIGARA[26], PIETRO VISCHIA[27,3], and NILOUFAR YAZDANPANAH[23,3]

[1]Lulea University of Technology, Lulea (Sweden)
[2]INFN - Sezione di Padova, Padova (Italy)
[3]Universal Scientific Education and Research Network
[4]Bush School of Government and Public Service, Texas A&M University College Station, Texas (USA)
[5]Faculty of Law, University of Trento, Trento (Italy)
[6]Soil Erosion and Degradation Research Group, Departament de Geografia. Universitat de València, Valencia (Spain)
[7]Australian Catholic University, Institute of Positive Psychology and Education, Sydney (Australia)
[8]Department of Industrial Engineering, Università degli Studi di Trento, Trento (Italy)
[9]Department of Neuroscience, College of Humanities, Arts, and Social Sciences (CHASS), University of California Riverside, (USA)
[10]Systems Neural Engineering Laboratory, University of Southern California Viterbi School of Engineering (USA)
[11]Chair for Scientific Computing, University of Kaiserslautern-Landau (RPTU), Kaiserslautern (Germany)
[12]Department of Psychology, University of Nevada, Reno, Nevada (USA)
[13]Institute for Better Health, Santa Rosa, California (USA)
[14]Department of Psychology, Philipps University Marburg, Marburg (Germany)
[15]Department of Psychology, Stockholm University, Stockholm (Sweden)
[16]Inria Center at the University of Bordeaux / LaBRI, Talence (France)
[17]Galician Centre for Mathematical Research and Technology (CITMAga) and Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela, Santiago de Compostela (Spain)
[18]ArcTech Learning Lab, Pervasive and Mobile Computing, Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Luleå (Sweden)
[19]Department of Neuroscience, Developmental and Regenerative Biology, University of Texas at San Antonio, San Antonio, Texas (USA)
[20]Department of Computer Science, University College London, London (United Kingdom)
[21]International Center for Tropical Agriculture (CIAT), Cali (Colombia)
[22]International Centre of Insect Physiology and Ecology (ICIPE), Nairobi (Kenya)
[23]Research Center for Immunodeficiencies, Children's Medical Center, Tehran University of Medical Sciences, Tehran (Iran)
[24]Constructor University, Bremen (Germany)
[25]Institute for Functional Intelligent Materials, National University of Singapore, Singapore (Singapore)
[26]Centre for Mind/Brain Sciences, University of Trento, Rovereto (Italy)
[27]Universidad de Oviedo and ICTEA, Oviedo (Spain)

Corresponding author: Mauro Da Lio (e-mail: mauro.dalio@ unitn.it).

ABSTRACT The recent rise in relevance and diffusion of Artificial Intelligence (AI)-based systems and the increasing number and power of applications of AI methods invites a profound reflection on the impact of these innovative systems on scientific research and society at large. The Universal Scientific Education and Research Network (USERN), an organization that promotes initiatives to support interdisciplinary science and education across borders and actively works to improve science policy, collects here the vision of its Advisory Board members, together with a selection of AI experts, to summarize how we see developments in this exciting technology impacting science and society in the foreseeable future. In this review, we first attempt to establish clear definitions of intelligence and consciousness, then provide an overview of AI's state of the art and its applications. A discussion of the implications, opportunities, and liabilities of the diffusion of AI for research in a few representative fields of science follows this. Finally, we address the potential risks of AI to modern society, suggest strategies for mitigating those risks, and present our conclusions and recommendations.

INDEX TERMS Artificial intelligence, scientific research, science ethics, computer science, physics, medicine, psychology, mathematics, geography, agriculture.

I. INTRODUCTION

A. THE FUTURE OF INTELLIGENCE

Thanks to a series of groundbreaking discoveries in exoplanetology, during the past thirty years we have gradually come to realize that planets similar to our own, where we may speculate that there be a significant possibility of emergence of carbon-based life forms such as those inhabiting Earth, are relatively common, with more than ten thousand candidates cataloged as of January 3, 2024[1]. Indeed, a handful of such candidates have already been identified within 120 light years of distance from us, and their number keeps growing as we improve our detection technologies. A recent estimate combining data from Kepler and Gaia missions using Bayesian inference [1] assesses in the range of 0.17 to 0.83 the number of planets in the habitable zone with masses between 1.0 and 1.75 Earth masses and orbital periods around stars of type F, G, or K between 237 and 500 days (see Fig. 1). Even though affected by significant uncertainties, these numbers imply that several billion earth-like planets exist in our galaxy alone.

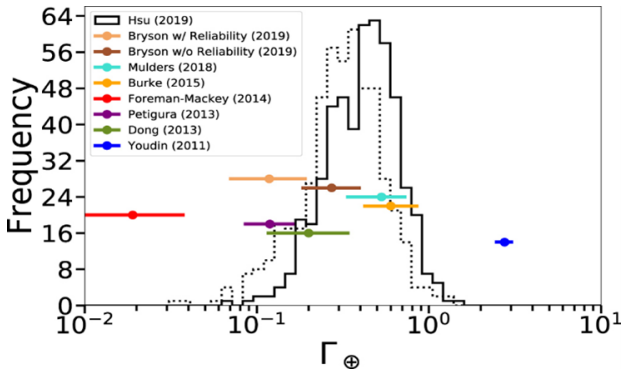

FIGURE 1. Inferred occurrence rate density ($\Gamma_\oplus$) of habitable-zone planets estimated by various studies. Displayed posterior densities for orbital periods of 237-500 days and different radius ranges refer to radii $Rp = 1 - 1.75\ R_\oplus$ (where $R_\oplus$ is the Earth radius) in solid black, and $Rp = 0.75 - 1.5\ R_\oplus$ in dotted black. Reprinted from [1].

On the other hand, geological studies indicate that planet Earth formed about 4.6 billion years ago [2], when it withstood gradual accretion through gravitational interactions from a disk of debris in orbit around our early Sun. Although less precisely, geological rock records indicate that only four billion years later did pluricellular life start to flourish on it [3], when a wealth of different creatures progressively inhabited its seas and later its lands. On such a time scale, the emergence of biological intelligence –which we here generically associate with biological life forms endowed with acquired self-awareness and entertaining communication and craftsmanship skills[2]– is a very recent phenomenon and might constitute only a brief parenthesis in the history of our planet. In light of the above considerations, the Fermi paradox (the contrast between the high probability of life emergence and the absence of its evidence, famously introduced by Enrico Fermi to question the hypothesis of widespread life in the universe) appears to have no objection to this line of reasoning.

When considering the universe from the point of view of its intelligence content, armed with estimates of the number of habitable Earth-like planets and knowledge of what happened on our planet until now, we are forced to assess what phenomena have the potential to cause mass extinctions. Many reasonably well-understood events of cataclysmic nature, from collision with asteroids and comets to solar flares, supervolcano eruptions, or nearby supernovae explosions, as well as slower evolutionary processes such as planetary motion instabilities, should then be assessed for their expected rates, which contribute to reduce the expected duration of our life span as a species, while also providing possible "restart" conditions to make the environment more suitable for life evolution. One must add several potential anthropogenic occurrences to those phenomena, including nuclear or biological warfare, climate change, and ecological collapse.

In the category of anthropogenic threats of relevance, an emerging one is the willful or serendipitous generation of artificial general intelligence (AGI), which develops goals of its own which are misaligned with the survival of humanity [4], [5]. A wealth of scientific fiction novels and movies have considered such a dystopian scenario, which, however, remains a highly speculative and hypothetical one, whose likelihood is impervious to estimation; we leave it aside here, as we will come back to it in more detail *infra* (Sec. V and Sec. VI). Yet even at the level at which we know it today, and in parallel to its offer of tremendous opportunities for progress, artificial intelligence (AI) is already a potential contributor to significant existential risks for humankind: for instance, its vicious use for the pollution of the information ecosystem with falsehood, undermining the basis of trust among individuals and nations and the concept/process of people electing their national leaders based on personal beliefs and preferences (since the opinion of people can be systematically biased using closed-loop AI systems and Internet-wide platforms). Various other consequences of the technological advancements in applied AI that have already taken place similarly constitute clear and present dangers today; we will examine them in the second half of this document (Sec. IV, Sec. V, and Sec. VI).

The probability of anthropogenic catastrophes is generally agreed to be much larger than that of natural extinction-level events and, therefore, contributes more significantly to reducing the expected life span of intelligent life on Earth, but it is harder to estimate. We can still summarize the situation by saying there is qualitative evidence that the rate of catastrophic phenomena is significant. This makes the lifetime of our species bounded from above.

Barring the possible yet arguably unlikely achievement of widespread colonization of nearby habitable planets by human beings, which would make intelligent life significantly more robust to global existential threats of both anthropogenic

[1] A comprehensive list is in https://exoplanets.nasa.gov/discovery/exoplanet-catalog/.

[2] A more detailed discussion of the meaning of the term "biological intelligence" is offered in Sec. II.

and natural origins, and using inductive reasoning in assuming that the history of our planet is not too uncommon, we must therefore come to terms with the idea that biological intelligence might be a comparatively rare occurrence in the universe: billions of Earth-like planets exist, yet the existence of intelligent life forms is likely only an ephemeral phenomenon on each of them. A different conclusion may instead be reached on AGI. In the past few decades, we have been lulled to perceive the rise in non-biological intelligence as a slow process by sitting on the initially linear-looking slope of what could soon manifest as an exponential curve. From the development of the theoretical underpinnings of AI in the first half of the twentieth century to the coming of age of powerful computers toward the end of the millennium, and then from the rise of smartphones to today's widespread AI systems, we were given the time to get accustomed to each new advancement without suffering a cultural shock. However, things are bound to change very soon, as the exponential trend in AI capabilities, diffusion, and overall impact on society are all becoming manifest. This has also revitalized discussions that experts have been having off and on for over sixty years about the possibility of creating AGI systems endowed with self-consciousness and general capabilities that would quickly transcend human intelligence, and it has also brought a wealth of arguments to those who argue about the likelihood of that scenario.

Although the topic is still controversial, the most poignant question today appears not if but when AGI systems will be produced. While opinions still differ widely, the majority view seems to be today that, whether it will be in 20 or 50 more years, AGI will arise on Earth: it has started to feel like an evolutionary necessity whose fuel is the enormous empowerment and profit it would guarantee to its creators and owners. And since AGI may be able to transcend most of the existential risks that biological intelligence is subjected to and is vastly more fit to endow itself with the means to become an interplanetary phenomenon, it is certainly not unreasonable to conclude –using again the hypothesis that Earth be a typical planet– that *the most common substrate of intelligence in the universe is artificial, and not biological*.

Despite its speculative nature, we deem the above observation highly significant, as it alone provides a substantial reason to view the future of artificial intelligence with heightened interest and concern: it is a phenomenon with a potentially transformative impact on different scales of space and time. This insight underscores the necessity for intensified scrutiny and engagement with AI developments. While we believe AI to be a logical and inevitable outcome of human evolution, how such a development unfolds remains partly within our control. Therefore, the consequences of our actions at this critical juncture are undeniably enormous.

B. RECENT IMPACTS OF ARTIFICIAL INTELLIGENCE

Over the past decade, we have witnessed many disruptive advancements in the capabilities of automated systems that display intelligent behavior[3]. In Sec. IV and V of this article we describe in detail the impact these systems have had on scientific research and our society, respectively. Here, to introduce the topic, we may single out the one among these developments that is perhaps the most surprising, besides being the most recent: the advent of large language models. In the matter of a couple of years, these systems have made a transition from being a topic of experimental research to settling as a centerpiece that fills an enormous void in the space of computer applications, effectively offering themselves as go-to oracles capable of providing detailed and relevant answers to almost any conceivable query. The impact that large language models are having in scientific research is highly significant, to the point that their assistance in the writing of scientific articles has turned from being a joke to becoming an acceptable, formally disciplined practice[4].

The acceleration that artificial intelligence has displayed over the past few years is most evident in market-driven applications, where it is powered by the enormous profits that the development of new AI systems may obtain; this is causing concern over the lack of relevant regulations. The situation is more controllable in scientific research due to two factors acting as regulatory brakes on applying new AI technologies. Firstly, progress is slowed by the absence of substantial funding sources that drive profit-oriented research. Secondly, many research areas that could withstand rapid acceleration and paradigm shifts, such as biology, medicine, and genetics, are under strict scrutiny by supervisory bodies. This results in the innovation rate aligning with these fields' characteristic approval cycles. The situation is, however, complex, as these bodies are not acting at a global level, and their power is limited. Under these circumstances, monitoring the situation by looking into the reality of public research across scientific disciplines appears crucial. Such a survey may inform us of better ways to handle the future challenges ahead of us. As a paradigm, we mention the recently issued EU Artificial Intelligence Act (from now on "AI Act"), the world's first comprehensive AI law [6], which starts addressing issues that will become more and more relevant in the coming years.

C. ABOUT THIS WORK

In this document, the production of which was organized by the Universal Scientific Education and Research Network (USERN)[5], we offer our views on the matter by discussing the implications of the current state of affairs in the development of AI for humankind, focusing in particular on scientific research, where we cover some of the most relevant areas of developments with our collective expertise as researchers. The purpose of this work is twofold. First, we wish to offer

[3]For a careful discussion and assessment of what can be qualified as such, see Sec. II

[4]*E.g.*, many journals today explicitly request authors to declare whether their work has been produced with the help of large language models—implying that the practice is acceptable as long as it is made manifest.

[5]The Universal Scientific Education and Research Network, https://usern.org.

a state-of-the-art survey on using AI tools for research. For this purpose, we examined a selected and representative set of scientific disciplines, considering the transformations already brought about by introducing AI techniques and forecasting their future evolution. Additionally, this text aims to spark a broad discussion on the opportunities and challenges posed by AI algorithms and systems in scientific research and human society by discussing their virtuous and nocuous or malignant uses. We do this with the long-term goal of bringing the topic to greater attention within the scientific community and keeping it firmly under scrutiny.

The article was written by surveying the advisory board of USERN, which comprises 600 scientists in 22 disciplines. The goal was to identify a team of experts who could provide an overview of the ongoing developments in AI and their impact on different disciplines (from a user perspective). The authors were asked to examine the challenges and opportunities that AI developments bring to their field of expertise and to summarize them in subsections of the manuscript. We then collaborated on these drafts to reach a consensus on the topics they covered. Special attention was given to identifying specific risks that AI developments may introduce in scientific research, as we believe we are well-positioned to make these assessments and that this is a valuable goal for our work.

We begin in Sec. II to establish definitions of intelligence and the related question of self-consciousness. In Sec. III we provide an overview of AI's state of the art and discuss potential future developments in several application areas. In Sec. IV we discuss the implications of AI for scientific research in a few selected fields. In Sec. V we consider the potential risks of AI to modern society, the strategies we might employ to mitigate those risks, and the benefits of its virtuous use for humanity and the environment. Finally, we present our conclusions and recommendations in Sec. VI.

II. DEFINITIONS OF INTELLIGENCE AND CONSCIOUSNESS

Before addressing the status and issues of artificial intelligence in science and society, we need to agree on what we mean by intelligence. The term is highly overloaded, as it may refer to a large number of different and independent abilities and competencies; indeed, its exact definition has kept generations of scholars entertained [7], including attempts at quantitative definitions indeed focusing on the extent to which an agent can be successful across a complex set of different competences [8].

A. INTELLIGENCE

For this work, we may agree that, in general, intelligence is the ability to solve problems. From a biological point of view, the environment poses various problems to organisms, from finding a nest to face a predator, from orienting in space to finding a mate [9]. Thus, we can expect the evolution of adaptive specializations in the realm of intelligent abilities. This is because the logical demands associated with different problems could imply a certain degree of functional incompatibility. Consider, *e.g.*, the logical demands related to spatial learning in food-storing birds as opposed to learning the features of the mother hen and siblings by imprinting. The latter must be limited to exposure to a specific period in life (critical period) and must be resistant to forgetting. Otherwise, the young animal would risk learning about inadequate objects or forgetting about the features of the mother and siblings [10], [11]. Food-storing birds, however, must not exhibit the same constraints in learning. Otherwise, there would be limitations to the places where food caching can be done depending on the period, and if memory would resist forever, it would prove impossible to erase the memory of the particular location of a food cache even when it has already been used, and it is now empty.

On the one hand, it is apparent that these functional incompatibilities, which are relative to the coherent constraints associated with a particular kind of problem more than to its implementation in a specific kind of substance, should be adhered to by any type of intelligence, biological or artificial. On the other hand, it is important to stress that there are also classes of problems for which we can expect no specialization but rather common and shared mechanisms [12]. As an analogy, artificial intelligence methods increasingly rely on massive, curated experiences, general pattern recognition algorithms, and modularized methods. Generalized adaptive processes are the case for associative, non-associative, and relational learning mechanisms, whose role is to allow an organism to figure out the causal structure of an environment in terms of either statistical regularities in the sequence of appearance of single events (habituation and sensitization), in the temporal relationship between two or more recurring events (Pavlovian and operant conditioning), or in the semantically and culturally attributed relations among events (multidimensional relational reasoning) [13]. Indeed, evidence suggests that biological organisms rely on similar neurobiological processes for these various forms of intelligent behavior. Leaving aside the question of the possible development of artificial intelligent systems from completely different substrates and with different bootstrapping mechanisms, one can reasonably assume that any form of artificial intelligence should also rely on similar and shared principles for non-associative, associative, and relational learning.

Considering the above understanding, Pei Wang's approach in his paper "On Defining Artificial Intelligence" [14] argues that intelligence, whether in humans or artificial systems, fundamentally entails adaptation with insufficient knowledge and resources. Wang stresses that the primary marker of intelligence is not just problem-solving *per se* but the ability to adapt to new and changing environments with limited information and capabilities. This perspective aligns with the biological examples provided, where different species evolve distinct cognitive strategies to handle their unique environmental challenges. Yet, all exhibit a form of adaptiveness that is central to intelligence.

Further, Wang proposes that any system exhibiting intelligence —whether organic or synthetic— must inherently oper-

ate under conditions of uncertainty and resource constraints. This notion resonates with the adaptive specializations observed in biological organisms, as each species' intelligence is shaped not just by the problems they solve but by how they optimize their cognitive processes within the limits of what is biologically feasible for them. Similarly, in artificial systems, this means designing algorithms that do not merely solve tasks but continuously learn and adjust to new data under the practical constraints of time and computational power.

Wang's conceptualization of intelligence also underscores the importance of a system's capacity to handle novel situations without pre-encoded solutions. This flexibility is crucial in biological and artificial contexts and is achieved through mechanisms that allow ongoing learning and adaptation rather than relying solely on hardcoded inflexible rules. Therefore, in creating artificial intelligence, it becomes imperative to develop systems that can dynamically learn from their environment in a manner analogous to biological systems, which continuously adapt through evolutionary pressures.

Thus, integrating Wang's theoretical framework into our understanding of intelligence may provide a more robust model that bridges the divide between biological and artificial systems. It suggests that the core of intelligence lies in the adaptive capacity to utilize limited resources and knowledge to navigate and optimize within a complex, changing environment. This holistic view can guide future research and development in artificial intelligence to focus not just on specific tasks but on creating systems capable of general adaptability and learning, reflecting the true essence of intelligent behavior observed in nature.

Moreover, considering the principles of evolutionary optimization processes in AI development, we see that biological and artificial systems traverse a space of emergent phenomena driven by fundamental interactions. Biological evolution, for instance, has led to the emergence of phenomenal experiences which we still do not fully understand. Similarly, computing substrates possess emergent properties that may arise from complex interactions within hardware and software components. These properties can be nearly orthogonal to the abstract logic processes typically utilized, leading to unexpected behaviors that are not directly programmed.

The limits of these emergent phenomena remain unknown, posing both opportunities and risks. Evolutionary optimization in AI could potentially design behaviors that emerge from interactions beyond deterministic information processing. These processes may exploit unknown properties of the computational substrate, leading to behaviors that are not predictable or controllable using traditional methods. Such emergent behaviors could result in forms of decision-making or problem-solving abilities that we cannot fully understand or anticipate, posing significant risks to safety, security, and ethical standards.

Indeed, AI systems' autonomy (and sometimes adaptiveness) are elements also included in legal texts. Most notably, both the AI Act, a regulation that has been approved but not yet entered into force, and the Organisation for Economic Cooperation and Development (OECD) [15] characterize AI systems as being machine-based systems that use inference to generate output from input and can influence their environment; they also show (and, in particular, according to article 3 of the EU AI Act, are actually "designed to operate with") varying degrees of autonomy and (possibly) adaptiveness after deployment. Both organizations consider autonomy as the ability to behave independently and operate without human intervention. At the same time, adaptiveness refers to the system's ability to continue evolving post-deployment due to its direct interactions with input and data.

The ability of an AI system to operate and modify its behavior independently of human instructions has become an object of attention by policymakers because of its ethical and legal implications, mostly revolving around human agency, autonomy, human oversight, accountability, and liability. Ultimately, if the core of intelligence lies in adapting to scarce information or capabilities, it is only proper that the regulation should account for these developments. Indeed, the (both moral and legal) responsibility for possible harmful effects on individuals or society, as well as the value-laden choice of which decision-making role to attribute to artificial intelligence systems [16] (including for which specific purposes and in which sectors), rests with humans [17].

B. CONSCIOUSNESS

Consciousness is a polysemic word, but here we shall consider it as «phenomenal experience» [18], *i.e.*, that it feels something to have mental states such as pain, seeing red, or smelling lavender [19], which is also usually referred to as the «hard problem» of consciousness [20]. So defined, awareness or "knowing" is a more evolutionarily continuous process based on the relative ability to respond to oneself and the internal and external environment and the regularities within and between them [21]. Various sensory and cognitive systems are relevant to awareness of that kind.

It is important to stress that not all cognitive activities need to be conscious. The issue of how and why humans and other organisms (or possibly artificial machines) do have consciousness cannot be equated with their having cognition; whatever sophisticated this cognition may be [22], [23]. There is plenty of evidence that advanced forms of cognition can be observed in the absence of consciousness [24], [25]. On the ethical side, note that we do not deny the possession of consciousness to human beings who, because of severe acquired or inherited disabilities, would appear unable of even the most elementary forms of cognitive activities [26]. Thus, the issue of consciousness in AI should be considered separately from evidence of how much intelligence there is in AI.

There is no agreement on the mechanisms underlying consciousness or the proper localization of its neural correlates. Some prominent theories, such as the Global Neuronal Workspace Theory (GNWT) [27] argue for a major participation of the parts of the brain that enable cognition, *i.e.*, the

frontal areas of the cortex, whereas others, such as the integrated information theory (IIT) [28], claim that consciousness would depend on brain areas involved in perception, *i.e.*, the more posterior areas of the cortex. These theories are sometimes called "front-of-the-brain" versus "back-of-the-brain" theories. However, neither of them adequately addresses the issue of phenomenal experience. GNWT maintains that a reduced subset of the information we constantly process unconsciously would be selected to pass through a bottleneck into a conscious "workspace;" there, the information would be integrated and broadcasted to other brain areas to make it globally available for decision-making and learning [29]. This theory does not address how conscious experience arises, which is mainly centered on cognitive access by attention. IIT, on the other hand, starts with five axioms about consciousness, namely that consciousness (1) is intrinsic to the entity who has it; (2) its composition is structured; (3) it is information-rich; (4) it is integrated rather than reducible to components; and (5) it is exclusive of other experiences. Then, the proponents of the IIT theory tried to develop mathematical descriptions to fit those axioms, but basically, the theory seems to deal with a measure of information rather than with phenomenal experience *per se*; criticisms have also been raised as to its mathematical parts; *e.g.*, see [30], [31].

Other scholars pointed to sub-cortical mechanisms for consciousness, in particular concerning pain [32]–[37] or about active movement with feedforward mechanisms mediated by efference copies to distinguish sensory stimulation impinging on an animal surface from that occurring by self-motion [38]–[42]. These latter hypotheses would agree with the possibility that consciousness could be observed in animals without a mammalian cortex, such as birds —see [43]–[45]; or even in organisms with quite different neural organization, such as insects or invertebrates in general [46], [47].

These kinds of theories appear of particular interest for AI, for they insist on the idea that to have consciousness, a body is needed, and that initial forms of consciousness should be represented by bodily reactions to stimulation (for instance, in the format of efference copies allowing distinction between "what is happening to me" and "what is happening out there"; see e.g., [41], [42]). If correct, these views suggest that embodiment (more specifically, the development of bodily reaction to sensory stimulation) would be more important than possessing symbolic cognitive abilities to develop an artificial consciousness.

1) Explanatory gap of the mind-body problem

The heart of the problem of consciousness is the issue of subjective experience and awareness, which is different from possession of cognitive abilities. It is difficult, if not impossible, to offer a non-circular definition between these three terms (subjective experience, awareness, and consciousness). Despite this difficulty, it is evident that conscious beings know what it feels like to be conscious. The subjective experience of what it is like to be in a particular consciousness state (which is referred to as the qualia of consciousness) cannot be reduced to the activity of the physical elements (networks of nerve cells in the brain) that enable it. This seemingly intractable mind-body problem is famously illustrated in Nagel's 1974 essay "What is it like to be a bat" [48]. Chalmers [20] later described this as the hard problem responsible for a gap between the physical world and subjective experience. No theories of consciousness have been able to address this so-called explanatory gap, and some do not even acknowledge it. As this applied to AI, however, a review of theories of consciousness becomes especially interesting when they imply evolutionary and experiential processes modeled in curated exposure to exemplars. Thus, theories of consciousness that fail to be specific about the kinds of evolutionary processes involved cannot readily be used by AI methods to address consciousness *per se*. We will review several theories before examining them in that light.

2) States of consciousness

It is widely accepted that consciousness is not a singular entity. Instead, it encompasses various states, particularly concerning experiential content, though not necessarily in the presence or absence of experience. These states can be categorized into two main types: global states and local states. Global states, or levels, pertain to an organism's subjective experience and are linked to arousal and behavioral responsiveness variations. Examples include wakefulness, dreaming, and sedation. On the other hand, local states pertain to specific conscious perceptions, emotions, or thoughts, often called conscious contents. These local states can be detailed at varying levels of specificity, from basic perceptual elements (*e.g.*, the color red) to objects (*e.g.*, a rose) or to comprehensive multimodal perceptions (*e.g.*, the beauty of a red rose).

3) Selfhood

One of the most remarkable features of the human species is the concept of the self and its ability to project and conceptualize itself in space and time. The self is a subset of local states of consciousness, encompassing the experience of selfhood, which includes emotions, intentions, volition, free will, body ownership, explicit autobiographical memory, and more. Self-consciousness, or the conscious awareness of the self, requires advanced cognitive abilities, including language, cultural understanding, cooperation, and empathy. The capacity to attribute mental states to others (known as the theory of mind) necessitates distinguishing between the self and the non-self.

The self is a highly complex construct. In his Principles of Psychology, William James (1890/1983) posited that self-awareness implies that the self acts as both object and subject, with an aspect of the self that knows (the knower) and an aspect that seeks to be known. The latter is called the Me-self, and the former the I-self. Aspects of the I-self include self-awareness, self-agency (ownership of one's thoughts and behaviors), self-coherence (perceiving oneself as a stable entity), and self-continuity (perceiving oneself as the same person over time). James identified aspects of the Me-self as

the material Me, the spiritual Me, and the social Me, noting that different social settings give rise to multiple social Me's.

Building on this foundation, we propose at least two aspects of the self: the core self, which is the relatively stable perception of one's persona, and the social self, which is more context-dependent and influenced by social role expectations [49]. This distinction is more than theoretical, as it provides new ways of understanding and treating emotional disorders associated with adaptations of the core self or the social self. As it applies to AI, if selfhood involves an acquired form of perspective-taking or meta-cognition, it may be essential to train large language models to respond to their problem-solving processes from a perspective or point of view rather than merely curating correct answers generated by humans.

4) Phenomenal and functional properties of consciousness

It is essential to distinguish between the phenomenal and functional properties of consciousness. The phenomenal properties refer to the experiential aspects of consciousness (the qualia or the subjective experience of what it is like to be conscious). In contrast, the functional properties relate to the role that conscious mental states play in an organism's cognitive processes. This functional aspect includes teleological functions (those shaped by evolution) and dispositional functions (the role a process plays within a larger system). For instance, consciously perceiving a red rose involves various potential functions, such as the ability to interact with it in multiple ways (*e.g.*, touching, smelling, picking, painting, or stepping on it). Depending on the context, this may lead to the formation of an episodic memory of the event and enable the generation of a verbal report about the experience. While these properties are definitionally helpful, they do not give clear guides on approaching the problem of consciousness in AI.

5) Global workspace theories

Global Workspace Theories (GWTs) are modeled on the blackboard architecture of AI, where the blackboard serves as a central resource for specialized processors to share and receive information. For instance, being conscious of a red rose signifies a relationship between my cognitive system and the object (the rose), which may be selected for further verbal or nonverbal processing after that. I might smell the rose and remark, "Look at that beautiful rose." The perception of the rose becomes conscious and globally accessible because the information can trigger a range of actions and thoughts.

The fundamental concept of GWTs is that sensory information gains access to consciousness when it is "broadcast" within an extensive neuronal workspace that spans higher-order cortical association areas, particularly emphasizing the prefrontal cortex [50]. This global workspace is accessible through a nonlinear network "ignition," where recurrent processing amplifies and sustains neuronal representations [51]. Once ignited, signals are amplified, allowing them to enter the workspace and thus become conscious.

Conscious mental states are globally available to a wide range of cognitive processes, including attention, evaluation, memory, and verbal report [27]. That way, conscious states guide behaviors and cognitions in a flexible, context-dependent way and are related to specific cognitive processes, especially attention and working memory. Consciousness and working memory are intimately related because attended working memory items similarly use the global workspace [50]. Once again, GWTs appear to help define the problem space but not to refine it with an eye toward AI methods.

6) Integrated information theory (IIT)

IIT proposes that consciousness should be understood in terms of a complex system's cause-effect power, resulting in integrated information. The degree of consciousness (and integrated information) can be quantified as phi (Φ), which measures how much information a system generates, compared with its parts independently. IIT assumes consciousness is widely distributed throughout nature, including in many non-biological systems, and might even occur in systems as simple as cell phones or computers [52], [53]. This approach seems to contradict philosophical arguments, including Searle's Chinese room thought experiment (see *infra*), and, in essence, "defines away" the problem of consciousness without solving it.

7) Higher order theories of consciousness

Global availability is just one aspect of consciousness [54]. Another key aspect is the cognitive system's self-referential capacity to monitor its processing, often called meta-cognition. Meta-cognition involves forming internal representations of one's knowledge and abilities. This higher-order level of processing differentiates it from first-order theories like GWT, leading to the term higher-order theories (HOTs).

HOTs propose that a mental state becomes conscious through these meta-representations, which target other representations. For example, the meta-representation "I am conscious of the red rose" refers to another representation (the rose), and this meta-representation constitutes conscious awareness. HOT has been applied to various domains, including visual experiences [55], emotional states like anxiety [56], and perceptual decisions [57]. HOTs emphasize anterior cortical regions, particularly the prefrontal cortex [58]. According to HOTs, some contents may remain non-conscious or unconscious if they are not the targets of appropriate meta-representational states. In contrast, other contents are necessarily conscious if accompanied by the right meta-representations.

HOTs rarely address global states of consciousness or the functions of consciousness [59]. Unlike many other theories, however, HOTs do appear to directly affect AI systems by suggesting that a second-order process of self-curation of exemplars may be key to generalized AI. In other words, it may not merely be the feedback of knowing when thinking is correct in its outcomes that is key; it is more being able

to self-generate that feedback from a consistent perspective or point of view and apply it to the process of the problem-solving itself. This kind of bi-phasic approach [21] has only recently begun to be modeled in AI learning systems, and it is not yet known if it will produce more rapid progress.

C. CONSCIOUSNESS IN HUMANS, ANIMALS, AND AI

1) Consciousness in animals

If we include the phenomenal aspect of the conscious qualia (*i.e.*, what it is like to be conscious) in the definition of consciousness, it is impossible to ascertain whether animals are conscious. Inferring the experiential state of an animal based on its overt behavior results in anthropomorphism. Indeed, animals, humans, and AI share aspects necessary for consciousness, such as perception, attention, memory, etc. However, none of these aspects and processes alone or in combination necessarily result in a conscious experience, as defined earlier.

As noted by Dehaene, Lau, and Koider [54], thirsty elephants can determine the location of the nearest water hole and move straight to it from a distance of up to 50 km [60]. This amazing feat combines several skills to integrate information, including smell, sight, memory, etc. Nest-building behaviors in birds are similarly complex tasks that may appear to a human observer as a thoughtfully planned and future-oriented behavior to achieve a specific goal. However, even the most complex future-oriented set of behaviors in animals can be explained through evolutionary selection and retention. We are not birds, and we cannot rule out the possibility that birds are consciously aware of their actions, worrying about their offspring, calculating the time it takes to complete the nest in relation to the shift in the season while looking forward to the task completion. We are humans, which is what humans would do and feel if we were birds. But we are not birds. And we are not elephants. We are humans.

The pronounced frontal cortex is an evident distinctive feature of the human brain, and there is good evidence to suggest that it supports the capacity for multimodal convergence and integration of cognitive processes. A comparison between humans and nonhuman primates identified a distinctly human component in the ventrolateral frontal pole and showed differences in interregional interactions within the ventral lateral frontal cortex in the two species [61]. There is also evidence to suggest that the complexity of the pyramidal cell phenotype in the prefrontal cortices is uniquely linked to human cognitive processing [62]. Additionally, humans possess circuits in the inferior prefrontal cortex for verbally formulating and reporting information to others. The relevance of such neurobiological data to consciousness is an object of active research, and progress in this area as it occurs may indirectly impact the issue of consciousness in AI systems.

Human language is often considered one of the most evident signs of conscious perception because information has reached this level of mental representation [54]. As is true for specific brain structures and circuits, although language may not be required for conscious perception and processing, the emergence of language may have resulted in a considerable increase in speed, ease, and flexibility. For example, damage to the primary visual cortex can lead to a phenomenon called blindsight. Patients with this neurological disorder report being blind in the affected visual field. Although they can localize visual stimuli in their blind field, they cannot report them.

2) AI and consciousness

Examining the role of artificial intelligence (AI) is helpful to clarify the boundary conditions of human consciousness. Consciousness is not reducible to an information processing system operating on formal symbols (a strong version of functionalism), as illustrated by John Searle's classic Chinese room thought experiment. Imagine an English-speaking person who does not know the Chinese language being alone in a room. In this room is an English instruction manual for processing any incoming information. The person then receives some Chinese characters through a slot in the door and is asked to process them according to the English instruction manual to produce other Chinese characters as output. Given enough time, the person will be able to make the output based on the instruction manual without understanding any of the content of the Chinese writing. Based on the generated output, an outside observer who knows Chinese will mistakenly assume a Chinese speaker is in the room.

Just like human operators in the Chinese room, computers have been able to use syntactic rules following instructions to manipulate symbols. Still, they arguably did not "understand" the meaning of those symbols. With that evidence, it might be concluded that the human mind is not simply a computer-like computational or information processing system, disproving what Searle termed strong AI. However, the recent development of deep learning, large language models, and, in general, of systems with super-human capabilities to organize, translate, interpret text, write computer code on demand, and other complex tasks puts those conclusions to severe testing. Notwithstanding, the above example illustrates the difference between two types of information processing computations: the selection of information for global broadcasting (which is accomplished by the English speaker in the Chinese room) and the self-monitoring of those computations. Only the latter is associated with a subjective experience (the qualia of consciousness); whether that is something we can generate with AI is a topic of the present debate.

III. STATE OF THE ART OF AI SYSTEMS IN FLAGSHIP TASKS AND PERSPECTIVES

In the previous section, we provided an overview of the complexity of defining consciousness, which is a central issue when considering artificial intelligence and the means we possess to understand its nature and limits. Recently, the emergence of Large Language Models (LLM) has brought the question of whether we can associate a form of consciousness with these systems. The answer at present is no, but given the fast development these systems are undergoing and the

plans –and a natural evolution of AI research focusing on LLMs– of enhancing them with sensory inputs and memory, the question will resurface soon, significantly hardened and complexified. In this section, we discuss these areas of AI development.

A. LARGE LANGUAGE MODELS, TEXT GENERATION AND PROCESSING

Understanding and producing natural language has been one of the primary goals of AI for several good reasons. Firstly, it is widely agreed that language is the hallmark that makes humans unique among all other animal species, and it is the means through which our most advanced intelligence is expressed [63]–[67]. Furthermore, the idea from which AI then emerged, the famous article by Alan Turing [68], proposed linguistic conversation as the best arena in which to exercise –and eventually accredit– intelligence to a computer. This is why recent neural models, capable of successfully processing and producing natural language with remarkable similarity to humans, have generated a significant sensation. The performance of these models was also surprising, considering that just a few years earlier, approaching human language proficiency seemed like an unattainable chimera. Furthermore, an even more advanced form of language -the ability to build mathematical models and make predictions about the future evolution of complex systems- is where the front line of AI research has now moved. Given the strong economic incentive for success in that area, it is not far-fetched to predict breakthroughs shortly. However, rather than considering future developments, we discuss the state of the art in language processing below.

1) Traditional natural language processing and artificial neural networks

The enterprise to process natural language with computers (Natural Language Processing, NLP) began as early as the 1950s, with attempts to translate text involving notable pioneers such as Warren Weaver and Yehoshua Bar-Hillel. Early software for translating from Russian to English, like GAT-SLC (Georgetown Automatic Translation - Simulated Linguistic Computer) and SYSTRAN (System of Translation), was developed [69]. These early attempts revealed the subtle complications that separate the sequence of symbols in a written text from its meaning. They highlighted the infeasibility of machine translation without a thorough understanding and formal framing of the countless complications of human language.

In the 1960s, an essential contribution in this direction was made through the mathematical treatment of syntax initiated by Noam Chomsky [70], [71]. His work enabled the automatic derivation of the syntactic structure of simple sentences [72]–[74] and early attempts at basic language understanding [75]. Over the subsequent half century, NLP made remarkable progress, expanding the modeling of natural language to multiple aspects such as proposition semantics [76], anaphora resolution [77], lexical semantics [78], and discourse representation [79]. However, integrating these pieces into a single system capable of processing language seemed impossible.

When artificial neural networks gained popularity in AI towards the end of the 1980s, intensive research developed on their use for natural language modeling [80]–[85]. The neural approach radically differs from NLP as it eschews Chomsky's theoretical framework and subsequent linguistic theories. The direction is radically empiricist, designing models that directly learn aspects of language from examples. The results aroused considerable interest in the cognitive and psychological field. However, the scope of these models was drastically limited to short, simple sentences and a reduced vocabulary, remaining far from natural language in its completeness.

A significant challenge in using artificial neural models for language processing stems from an apparent irreconcilable discrepancy between the two formats. Language is an ordered sequence of auditory signals or written symbols, while a neural layer is a real vector with a fixed dimension. Designing a transformation from words to numerical vectors that preserves aspects of the words' meaning is problematic. Representing words becomes more challenging when transitioning from single-word morphology to syntax. Moreover, feed-forward neural networks are static, making establishing a sense of ordering for multiple words in a sentence far from straightforward.

Various coding strategies were proposed for solving the representation issue, while the ordering issue was addressed using recurrent networks. However, none of these solutions was effective when moving from a small controlled language to complete language processing. Recurrent neural networks struggle to maintain relevance for words that are too distantly placed yet syntactically related. A slight improvement was made with Long Short-Term Memory (LSTM) [86] and Gated Recurrent Unit (GRU) [87].

On the other hand, neural networks are comfortable dealing with continuous signals. With the advent of deep learning, tasks like text-to-speech and speech-to-text conversion have been essentially solved [88], [89]. Systems like Google's Tacotron and Facebook's Wav2Vec have dramatically improved the fidelity and understandability of generated speech and transcriptions.

2) The Transformer architecture

The Transformer architecture, invented by the Google team [90], combines two effective strategies that address the crucial issues noted in the previous section. The first is the word embedding technique, introduced by Mikolov *et al.* [91], which learns from examples the optimal mapping from words to vectors of neural activity. Its primary feature is that the vectorial representation is meaningful, allowing for manipulation and yielding results consistent with aspects of lexical semantics.

The second strategy is a mechanism called "attention" [92]. This technique dynamically identifies relevant information and relationships among words in a sentence. The Transformer employs these strategies innovatively: on the

one hand, word embedding is learned while the entire neural model absorbs everything from corpora; on the other hand, the attention mechanism entirely replaces recursion, presenting all words along with their vectorial embedding simultaneously as input.

The Transformer architecture retains elements reminiscent of the autoencoder (see Sec. III), featuring both encoding and decoding components. Its seemingly straightforward task involves reproducing the sequence of words it encounters as input. The original version was designed for translation, with an encoder for the input text and a decoder for the text generated in a different language. A simplification was later adopted by GPT (Generative Pre-trained Transformer), consisting only of a decoder part, primarily for generating text by completing a given prompt [93]. The famous public interface ChatGPT is based on later models of the GPT family [94].

The primary innovation of the Transformer lies in a simple heuristic for gauging the relevance of each word in the sequence relative to the current context, leveraging four matrices known as key, query, value, and output. The key matrix transforms the incoming words within the sequence, while the query matrix transforms the words generated in the output. The scalar product of these two matrices modulates the outcomes of the other two matrices. The value matrix operates on the input sequence, while the output matrix affects the output sequence. This mechanism governs which information is extracted from the source words and how it is incorporated into the destination. The matrices are learned dynamically during the training process, in tandem with the conventional feed-forward neural weights of both the encoders and decoders. Moreover, the representation of words as neural vectors is not fixed from the outset but is gradually acquired through exposure to extensive corpora. In a discursive manner, the attention mechanism produces a vector where information from all the words preceding the current one is mixed, weighted according to how relevant the previous word is to the current one. Here, the synergy with the other fundamental mechanism of the Transformer comes into play: word embedding. The efficiency and completeness in capturing every relevant word information in a numerical vector, in every possible context of use, allow relying on simple operations of linear algebra to capture the meaning of the reciprocal syntactic and semantic relationships in a text.

In addition, the entire token expressed as an embedded vector is divided into H portions, called heads, and the identical mechanism described is applied separately to each head. Only in the end are the various portions rejoined. The idea is that an embedded vector combines different properties of a word, and specific categories (*e.g.*, the tense of verbs or the gender and number of nouns and adjectives) always occupy the same portions of the vector. Therefore, it is convenient to process the network of relationships separately between the characteristics of the various words in the text.

Neural language models powered by the Transformer architecture can produce extensive essays comprising thousands of words, conveyed in well-articulated language and enriched with substantial content. The sudden and unexpected appearance of models capable of fulfilling Turing's dream – conversing with human beings [95]– in addition to opening the doors to many practical applications, has sparked an intense theoretical debate. Questions are being raised about how much neural language models truly understand [17], [96]–[99]; what the appropriate methods are for investigating their nature and their potential [100]–[103]; whether they possess the ability to understand the minds of their users [104]–[106]; and even whether they possess consciousness [107], [108].

While AI has made significant strides in understanding and generating natural language, there remains ample scope for improvements. The journey ahead necessitates technological innovations and an intense focus on the ethical, transparent, and responsible use of these systems. The future that combines these will unfold a new era of AI language processing capabilities for a breadth of practical cases.

B. IMAGE PERCEPTION AND GENERATION

Artificial vision is the field where deep artificial neural networks have unexpectedly revolutionized AI, approaching human performance for the first time in history [109]. This revolution began just over a decade ago [110]. In the wake of this revolution, a new and unique perspective has emerged: the combination of visual perception and image generation. This concept was not present in the 50 years of image processing tradition that preceded deep learning, which primarily meant recognizing its content when processing an image. In that era, image synthesis was a separate domain entirely [111].

The idea that perception and generation are unified in vision is supported by fascinating cognitive science and neuroscience theories and evidence. However, these theories and evidence have rarely been explicitly used to inform the development of artificial vision systems. One immediate example of this unity is the phenomenon of visual imagery. When a person tries to picture an object or a situation, they are essentially re-enacting an internally simulated perception of that object or situation. It was debated whether visual imagery shares features and mechanisms with visual perception for a long time. While Stephen Kosslyn was a supporter of this hypothesis [112], Zenon Pylyshyn flatly rejected it [113]. Recent evidence suggests that Pylyshyn was wrong. Studies have shown that, in the absence of the visual stimulus, the primary visual cortex is activated during visual imagery, albeit with a simplified representation of the object of interest [114]. This infers that visual imagery is not simply a form of imagination but rather an internal simulation based on the exact neural mechanisms of visual perception. More recent evidence suggests a significant overlap in neural processing during perception and imagery, encompassing the visual, parietal, and frontal cortices. This indicates that the two processes are more closely related than previously thought [115]–[117]. Reusing the same neural circuit is the foundation of mirroring, mental simulation (thinking), imitation, deliberation, and mind reading [118].

1) The idea of autoencoder

There are several ways to implement the unified approach of perception and generation in deep artificial neural networks. The oldest, yet still a fundamental component in the most advanced models, is the autoencoder, a neural network whose task is to reproduce its input as output by learning a low-dimension representation of the input. This may seem like a straightforward objective, but as the autoencoder learns to replicate its input, it demonstrates the ability to construct highly insightful and concise internal representations of the data. This concept has been around for quite some time [119] but has more recently served as the cornerstone for the transition from shallow to deep neural architectures [120], [121]. The critical challenge of training neural architectures with multiple internal layers was initially addressed by associating each layer with a Restricted Boltzmann Machine [120]. This allowed the layers to be individually pre-trained in an unsupervised manner. Adopting autoencoders overcame the training costs associated with Boltzmann Machines. Autoencoders permit the training of the whole network like an ordinary fully connected layer. The key idea here is to utilize the same input as the target output, which trains each layer to optimize input reconstruction. The overall result is a regularization of the entire model, akin to the one achieved with Boltzmann Machines [122].

Regardless of specific implementation details, an autoencoder's fundamental structure consists of two neural models. The first is the encoder, responsible for computing a compact representation of a high-dimensional input. The second is the decoder, often called the generative model, which generates high-dimensional data using the low-dimensional compact representation as input. In the initial implementation of the autoencoder for handwritten digit recognition [120], the encoder was constructed by stacking feed-forward layers with a decreasing number of units. At the same time, the decoder consisted of feed-forward layers with a number of units that mirrored those of the encoder in reverse order. In subsequent developments, the encoder typically comprises a hierarchy of convolutional filters interspersed with nonlinear down-sampling. This architecture essentially follows the design introduced by Hinton [110], which is derived from the earlier Neocognitron model proposed by Fukushima [123]. The corresponding decoder component adopts the deconvolution approach [124]–[126], which alternates convolution filtering with unpooling, ultimately restoring the high dimensionality of the input image.

2) From autoencoders to diffusion models

The autoencoder stands as a fundamental concept bridging the realms of perception and generation, and it is one of the few outcomes of AI with a robust neurophysiological counterpart [127]. The autoencoder has served as the genesis of numerous divergent approaches, departing from their initial biological inspiration and evolving into the cornerstone of today's most influential AI generative models for image generation.

While the primary focus of deep learning developments at the start of the past decade was on enhancing the successful Convolutional Neural Networks (CNNs) architectures, the autoencoder found its path, mainly due to its compelling self-supervision potential. One notable achievement in recognition systems inspired by the autoencoder paradigm is the U-Net, initially designed for medical applications [128]. Kulkarni [129] harnessed the latent autoencoder representations effectively for discriminating individual graphical attributes within images, including location, pose, lighting, texture, and shape. During this same period, a probabilistic variant of the autoencoder gained rapid popularity, wherein each element of the latent vector incorporates both the mean and variance parameters of a Gaussian distribution. This variant is commonly referred to as the variational autoencoder [130], [131], and, akin to the deterministic autoencoder, it offers the valuable advantage of self-supervision [132]. A notable image processing model that combines variational autoencoders with convolution layers is PixelVAE [133].

A pivotal advancement that greatly influenced the exclusive generative utilization of autoencoders was the introduction of stacked denoising autoencoders [121]. In this innovative approach, the original image undergoes a gradual degradation by adding Gaussian noise. Each autoencoder stage takes a specific level of degradation as input, aiming to encode the immediate level with reduced noise in its output. Through the sequential chaining of these autoencoders, there is a progression from an almost uniformly noisy state to the best achievable reconstruction of the original image.

The immediate benefit of this framework in typical recognition or detection tasks is its effectiveness in handling noisy images. However, the true novelty lies in the framework's ability to generate high-quality images by randomly sampling a matrix filled with Gaussian noise. Denoising autoencoders share similarities with formulations inspired by principles from physics [134], where degradation occurs through probabilistic diffusion, resembling processes in non-equilibrium statistical physics [135]. The degradation operation may be produced by down-sampling, blurring, or Gaussian noise [136]. In all instances, the reverse transformation is learned through conventional neural approaches, and these methodologies are collectively referred to as diffusion models [137].

One of the early and still widely used proprietary tools for generating images with diffusion models is Midjourney[6]. In a parallel vein of development, a team at Ludwig Maximilian University of Munich [138] shifted the perturbation process from the input/output of the autoencoders into the latent space. This not only reduced complexity but also notably enhanced image quality. They called this solution stable diffusion, with support from the company Stability-AI.

Thus far, the intersection of perception and generation has yielded formidable methods for synthesizing images without any initial visual cues; they can be generated entirely randomly (see Fig. 2 for a couple of examples). However,

[6]https://updates.midjourney.com/.

FIGURE 2. Images generated by DALL-E 3 when giving as prompt the title of this paper: "Artificial Intelligence in Science and Society: the Vision of the Universal Scientific Education and Research Network."

imagination is inherently non-random, and for most practical purposes, a trade-off is necessary between creativity and alignment with the intended image's purpose. This raises the question of how to "condition" or "guide" the model toward generating the desired type of image.

With the wide availability of deep learning classification models, it has become easy to integrate generative models with classifiers. This allows for selecting images with a high probability of being categorized into a specific class. Consequently, users can designate one or more standard ImageNet class labels to steer the image generation process [139].

3) The Transformer milestone

The Transformer architecture has also affected the field of computer vision, providing an excellent solution to the previously mentioned challenge: guiding image generation according to user preferences. Whether you are a fashion designer, a magazine cover illustrator, or an altermodern painter, there are no more effective means of articulating your creative vision than through natural language. The emergence of neural language models equipped to comprehend natural language descriptions reveals an exciting new avenue for exerting precise and captivating control over image generation.

The OpenAI team achieved a groundbreaking milestone by introducing the first Transformer-based image generation model, known as DALL-E [140]. This remarkable achievement was made possible through a series of research endeavors that, driven by the success of Transformers in natural language processing, explored the direct application of this technology to image processing [141]. A pivotal aspect of this development is the concept of "image tokenization," which involves efficiently segmenting images into a one-dimensional sequence of encoded vectors. While preserving all visual features captured by CNNs in a linear token sequence can be challenging, it proves sufficient when the objective is to align images with natural language descriptions to control image generation. In the case of DALL-E, this challenge is addressed by preprocessing images with variational autoencoders, compressing a 256x256 RGB image into a 32x32 grid of tokens, each having 8192 potential values. The text component is processed during generation to generate the most likely sequence of "image tokens" as the output. In the second version of DALL-E, the image generation aspect is further enhanced by including a diffusion model decoder, which derives its latent information from the Transformer-tokenized representation of the predicted image. This combination of a Transformer-based approach for simultaneous learning of text and related images, coupled with a diffusion model to enhance output quality and resolution, has become a pivotal factor in the ongoing advancement of generative models. Notably, Google's ImaGen Google [142] is a testament to the continued evolution of such models. Meta has made another notable achievement with the DINO model [143], a self-supervised learning algorithm for training Vision Transformers (ViTs) without needing large amounts of labeled data. DINO works by training two ViTs, a student and a teacher. The student is trained on augmented or distorted versions of images, while the teacher is trained on the original images. The student model is then asked to predict the original image from the augmented image. By doing this, the student model learns to identify and preserve the images' essential features, even when distorted.

Generative AI rapidly integrated itself into the creative toolkit of artists and designers. A pioneering moment in the art world occurred when Jason Allen's piece, "Théatre D'opéra Spatial," generated using the Midjourney tool, obtained the first artistic award ever won by an AI-created work at the Colorado State Fair Fine Arts Competition. As of September 2022, OpenAI reported that over 3,000 artists use their DALL-E AI system from 118 countries.

Unsurprisingly, the art world did not uniformly embrace this innovation with enthusiasm. Many scholars and critics raised concerns regarding potential risks to the integrity of an artist's work and the evolving concept of art and creativity. Consequently, a profound and ongoing debate has emerged, addressing the implications of generative AI from a multitude of perspectives: psychological [144], philosophical [145], and sociological [146]. This debate is intricate and multifaceted, a topic that is beyond the scope and purpose of this discussion. Instead, we believe that the sentiments of media researcher and artist Joanna Zylinska [147] eloquently capture the essence of this discourse: "It is worth remembering that the anxiety evoked by ML technology in relation to art has historical precedence. In the early 1820s, for example, it was feared that the invention of photography would lead to the death of painting. Instead, photography generated an explosion of new ways to see and create images–including painted ones".

Equally unsurprising is the regulatory world's struggle to deal with the disruptive effects of these technologies in various legal domains, as they impact multiple areas, such as privacy, copyright, civil liability, cybercrime, and disinformation. While it will take some time for these various sectors to accommodate Generative AI's specificities, the AI Act proposes harmonized rules for the design, development, and post-market monitoring of AI systems throughout the European Union's market. It uses a risk-based approach, categorizing artificial intelligence systems into unacceptable, high,

limited, and minimal risk levels. On this basis, the regulation offers a three-pronged approach concerning Generative AI.

First, generative models will be categorized according to the risk levels identified for the AI systems into which they are implemented. Thus, they could, for instance, be prohibited (unacceptable risk) or subject to requirements before entering the market (high risk).

Second, the limited risk category is particularly relevant for systems implementing Generative AI models, and notably, its requirements can be combined with those of high-risk systems. Article 50 establishes that artificially generated or manipulated content (text, audio, image, or video) must be marked as such by the provider in a machine-readable format (*e.g.*, via watermarks). The deployer must disclose the artificial nature of the output. At the same time, humans must be informed that they are interacting with systems such as chatbots (typically based on language models). These so-called transparency measures for limited risk systems relate to the nature of the system and its output. They are intended to promote trust, prevent manipulation and disinformation, and enable informed choices by individuals who interact directly with an AI system or are exposed to AI-generated content as a kind of informed consent [148].

Third, general-purpose AI models, including Generative AI models, will also be subject to specific documentation and copyright compliance requirements (article 53). In addition, general-purpose AI models with systemic risk will be required to assess and mitigate potential systemic risks, ensure adequate cybersecurity, and document and report serious incidents and relative mitigation actions (article 55).

C. AUTONOMOUS VEHICLES

Autonomous vehicles (AVs) are one of the simplest forms of embodied AI. They have a narrow AI goal, which is safe and efficient locomotion on roads (safe and efficient driving). The vision of self-driving vehicles has notable precursors in the European "Prometheus" project from the early 1990s and the US DARPA Challenges (2005-2007). As a form of embodied intelligence, AVs have a body that must respond to environmental stimuli while pursuing long-term navigation goals. Given the definitions we have provided in Sec. II, they are examples of agents that require cognitive abilities [149] (and are potentially suitable for consciousness, according to what we posited in Sec. II).

The dominant sensorimotor architecture (Fig. 3, top) of today's AVs follows a traditional control engineering approach consisting of three sequential functional blocks [150]: 1) a perception block that creates a symbolic "representation" of the road environment, 2) a decision block that plans a behavior by reasoning over the symbol set, and 3) a control block that actuates the behavior through lower-level control loops.

The definition of the atomic symbols for the output of the perception layer is made by human designers. Thus, although the perception layer may contain large trainable neural networks that classify sensor data into different entities, the

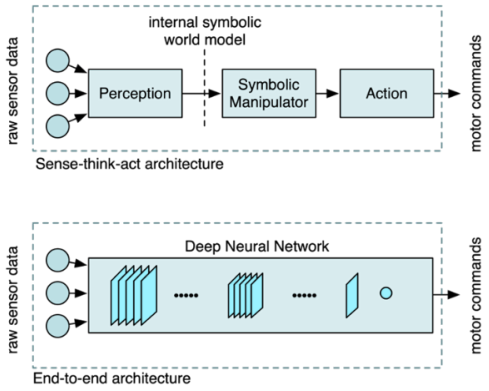

FIGURE 3. Top: the sense-think-act architecture used for self-driving systems. Bottom: the end-to-end architecture. Both are functions that map sensory data to control actions. The systems will always respond in the same way to the same input (except for noise) and cannot be considered as having conscious internal initiatives.

output classes are predefined and fixed. If an entity that does not belong to the predefined classes occurs, it is forced into the closest class or ignored. That was the source of a fatal accident involving a UBER car that hit a pedestrian who was walking with a bicycle, as such a pattern could not be recognized into a predefined class by the vehicle's perception system [151], [152]. A review of AI techniques for crash prediction may be found in [153].

Then, following the choice of the symbols for the world "representation," engineers design the algorithms for the decision block. From an engineering perspective, this is expected to guarantee that the systems are verifiable and that their safety and performance can be guaranteed. Unfortunately, the complexity of the situations that a self-driving agent may face generates complex rules and algorithms –see, *e.g.*, [154], [155].

The two shortcomings above may explain why the recent development of autonomous driving technology is impeded by the continued emergence of many edge and corner cases [156]–[159], with a long tail of unexpected situations in which cars are unable to act [160] or act dangerously [151], [152].

From the perspective of AI, self-driving vehicles engineered with the sense-think-act architecture look closer to complex control systems rather than genuine artificial intelligent agents. They are programmed by human beings and evolve only with human re-coding. As such, these systems are deterministic controllers. Their output depends on the input and may vary only because of noise in the perception-action chain. One can consider these systems to possess some cognitive abilities. Still, they are not conscious, have no "free will," and will not have any self-initiated goal other than responding in a programmed way to the environment and vehicle states.

As an alternative to the sense-think-act architecture (Fig. 3, bottom), NVIDIA demonstrated a lane-keeping function

trained end-to-end, given raw sensory input and steering examples from an expert human driver [161]. However, in the following studies, Waymo stated that "simple imitation of a large number of expert demonstrations is not enough to create a capable and reliable self-driving technology" and studied a perturbative approach to creating synthetic data that help to generalize self-driving abilities [162]. These efforts witness the attempt to explore different cognitive architectures, which are discussed next. Although the implementation is different, the following systems are still simple functions that map input to output and are unconscious.

D. ROBOTICS AND HUMAN-ROBOT COOPERATION

Compared to autonomous vehicles, robots are more complex forms of embodied AI. They may have a more comprehensive set of AI goals involving abilities such as manipulating objects (for example, the dexterous manipulation of tissues is exceptionally challenging), locomotion in completely unstructured environments, and interaction and cooperation with humans.

A significant difference between embodied intelligence and LLM is that embodied robots can act on the environment and observe the consequences of their actions. In this way, they can explore the world in autonomy via generalized motor babbling and construct predictive models of increasing abstraction levels. This gives embodied AI the ability to refine behaviors, acquire new behaviors, and, in principle, new goals. If a robot develops selfish goals and behavior, the gap between a cognitive system and consciousness shrinks, albeit it is hard to say whether this mechanism alone may be sufficient to reach consciousness. Below is an example of a robot capable of extending its sensorimotor system. However, let us first recall precursor cognitive architectures.

The subsumption architecture [163] generates behaviors using perception-action loops organized hierarchically. It starts from simple behaviors at the bottom and creates more complex behaviors at higher levels by subsuming lower levels. The topic of action selection, *i.e.*, which of the many behaviors created in parallel by subsumption architectures takes control of the agent, was studied by Prescott [164], indicating that a centralized switching mechanism is preferred. With centralized action selection, the system is scalable: new behaviors can be added to the stack (or learned). As long as they provide the behavior reward on a standard currency scale, they become immediately available for selection to the agent. Adaptive behaviors and learning of new behaviors are obtained seamlessly with layered control architectures (the name given to subsumption architectures with centralized action selection).

Notably, layered control architecture in robotics parallels the affordance competition hypothesis in natural cognition [165], where the brain's "dorsal stream" responds to affordances by instantiating potential actions in parallel. Then, the basal ganglia selects one using a competition process based on estimating their rewards.

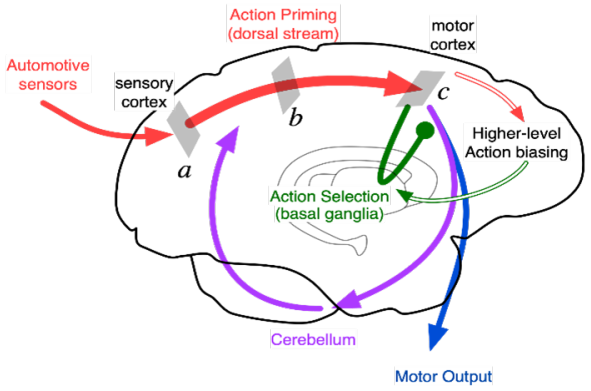

FIGURE 4. A cognitive architecture that can learn models of the world and use them to self-extend its sensorimotor repertoire.

Layered control architectures have two primary learning loci: one is learning biases in the action selection process, which can be used to steer the behaviors according to long-term goals [166]. This process may be realized through reinforcement learning or other methods. The second learning mechanism consists of expanding the subsumption stack, which can happen in several ways. A promising approach is generative AI: the perception-action loops are made of generalized autoencoders with various branches, which resembles the organization in Convergence Divergence Zones proposed by Damasio [167]. With generative AI, different objects are encoded separately in the latent space, avoiding using predefined symbols. The symbols are instead learned. These symbols are often qualified as "modal" because they are linked to the states of the neural convergent and divergent zones that produce the particular patterns in the latent space. Re-activation of those states can reconstruct the corresponding input or predict corresponding motor behaviors (similar to image perception and generation but declined to sensorimotor control). Learning the evolution of modal symbols across time corresponds to learning predictive models. In turn, predictive models can be used to imagine hypothetical events and derive novel action strategies [127], [168].

A final benefit of the layered control architecture is emergent human-robot interaction [169]. This relies on a modified version of action selection where actions are selected according to the similarity with the activity of a human with which the agent intends to cooperate.

An example of a robot with the above capabilities is the self-driving agent developed in the EU-funded project Dreams4Cars[7]. The robot sensorimotor architecture is shown in Fig. 4, superimposed on the outline of a brain, inspired by Cisek's affordance competition hypothesis. The red arrow is the primary perception-to-action loop. It creates many potential actions by responding to affordable actions offered by the environment. The action selection loop selects one of them via a competition process and gates it to the motor system.

[7]www.dreams4cars.eu

In this way, the agent achieves basic adaptive behavior [170]. However, by noting the consequences of actions, the agent can learn models of the world (the violet feedback loop and the red arrow in the inverse direction). These models have various uses [118], [171]. Among them, learned models of the world can be used to simulate the real world mentally (without engaging in dangerous actions) to develop action strategies safely. Once created, these new sensorimotor behaviors can be incorporated into the primary action priming loop and extend the agent's capabilities [172], [173]. Alternatively, biasing strategies can be developed to steer action choices towards those more beneficial for the agent [174] (this is where hypothetical consciousness might control the agent).

Concerning the purpose of the present paper, we note that, compared to the self-driving agents of Fig. 3, several feedback loops now allow the agent to self-extend its capabilities. Hence, while at any given time, an agent always responds in the same way to external stimuli (hence, it is not conscious), with time, it can evolve its abilities and improve its ways of dealing with problems. The agent evolution was supervised in Dreams4Cars by human operators who decided which scenarios should have been analyzed. However, in principle, an agent could explore hypothetical scenarios randomly and –without supervision– we cannot exclude that the agent develops selfish (but dangerous for humans) behaviors. Hence, some mechanism for approval of learned behaviors should be set. Compared to humans, artificial agents can also share learned behavior in a much easier way. This may be useful (for example, a population of cars can learn a safer behavior from one rare, dangerous event that happened to one agent). Still, it can also create the risk of quickly spreading undesirable behavior.

Thus, from the perspective of artificial intelligence, the robotics cognitive architectures described above are capable of different forms of learning. Hence, they can evolve autonomously. The agents have cognitive states. However, given the definition we provided in Sec. II, the simple possession of cognition does not imply such agents are conscious. Nonetheless, guaranteeing ethical and safe self-evolution is an open point.

E. MULTIMODAL SYSTEMS AND AGI

In recent years, multimodal systems have emerged as a critical area of focus within AI research. These systems concurrently utilize multiple input or data types (*e.g.*, text, image, video, audio) to make more nuanced interpretations, decisions, and predictions. Key advancements include image captioning, visual question answering, and audio-visual speech recognition.

Various architectures have been developed to tackle multimodal tasks as part of the continual innovations in AI. We mention a few of them below.

- Multimodal Transformers. Recent technologies have seen an extension of transformer-based models from the realm of Natural Language Processing (NLP) into the multimodal domain. These models are proficient at dealing with multimodal data using self-attention mechanisms that enhance their learning capabilities from multiple data sources.
- The Late Fusion CNN-RNN model is a hybrid architecture for video classification tasks. The model employs a CNN for extracting frame features and a Recurrent Neural Network (RNN) for sequentially encoding those features. The model's architecture allows it to process videos in a manner that accounts for both the visual content (CNN) and the temporal dynamics (RNN).
- ViLBERT (Vision-and-Language BERT) is a model architecture designed for tasks that require understanding both visual and textual content. It features two parallel BERT (Bidirectional Encoder-Representations from Transformers) models, one each for vision and language, connected through co-attentional transformer layers. This architecture allows effective communication between the textual and visual streams.
- Developed by Facebook AI, MultiModal Framework (MMF) is a multimodal research framework to facilitate the development of vision and language models. It is used in models like VisualBERT and MMBT (Modular Multimodal BERT), which effectively combine text and image data for various tasks such as image-text matching and visual question answering.
- LXMERT (Learning Cross-Modality Encoder Representations from Transformers) is another prominent example of multimodal transformer models. It accommodates visual and textual data, possessing separate encoders for each and a cross-modality encoder for integrated processing.

The architectures mentioned above underline the potential held by multimodal systems in advancing AI's capabilities. However, despite these advances, finding an effective way to synchronize data from different modalities continues to be challenging, driving ongoing innovations in this area.

1) Limitations and improvements

Despite the progress shown by the above systems, challenges persist. For instance, integrating heterogeneous data and synchronizing different modalities remain significant hurdles. While the field has seen impressive growth, there is still much room for improvement to advance models' accuracy, context understanding, and real-time performance in complex, real-world environments.

2) Artificial General Intelligence

The ultimate goal in AI research is to achieve artificial general intelligence, an AI system with generalized human cognitive abilities to understand, learn, and apply knowledge across a broad range of tasks. Despite substantial progress in the past few years, AGI remains hypothetical, with tangible examples yet to surface in the real world. While multimodal systems are a steppingstone towards developing AGI, it is critical to note that AGI's realization involves addressing challenges beyond technical feasibility, including ethical implications, societal

readiness, and regulatory landscape adaptation. Wreturnck to these issues in Sec. V.

3) The non-axiomatic reasoning system (NARS)

NARS, or the Non-Axiomatic Reasoning System, is a model designed to embody a theoretical understanding of intelligence that is abstract enough to apply to human cognitive processes and artificial intelligence systems. This model is distinctive because it does not focus on specific application problems but aims to provide a general-purpose system that can handle various tasks, including those not anticipated by the system or its designers [175], [176].

NARS is intended to be an Artificial General Intelligence (AGI) system. It is designed to abstract intelligence from human intelligence, focusing on the ability to adapt and function in a wide range of situations, not limited by the specific biological characteristics of human brains. The system is structured to handle tasks and problems that are not predefined or anticipated, making it a general purpose. This aligns with the goals of AGI, which aims to create systems that can perform any intellectual task that a human being can, rather than being confined to narrow, specialized tasks typical of many contemporary AI systems.

The architecture of NARS emphasizes adaptability, resilience to insufficient knowledge and resources, and the capability to learn from experience, all of which are key attributes desired in AGI systems. It does not attempt to model human thought processes or learning methods directly; instead, it develops its problem-solving strategies based on its interactions and experiences, further underlining its alignment with the broader objectives of AGI.

The core principle behind NARS is the "Assumption of Insufficient Knowledge and Resources" (AIKR), which posits that any intelligent system must operate under limited knowledge and constrained resources. This leads to a system that, rather than striving for perfect or optimal solutions, prioritizes flexibility, adaptability, and originality. NARS uses a unique approach to reasoning based on experience and adapts over time, reflecting the system's ongoing interaction with its environment. This model contrasts with traditional AI models that often rely on fixed, pre-programmed knowledge or purely statistical learning mechanisms [177].

In summary, NARS represents a significant departure from conventional artificial intelligence paradigms by emphasizing an experience-based and adaptive reasoning process. Its flexibility may be the key to avoiding many potentially dangerous behaviors that an autonomous agent endowed with superintelligence might otherwise develop.

F. BRAIN-COMPUTER INTERFACES

Brain-Computer Interfaces (BCI—also known as Brain-Machine Interfaces) are systems that can translate observed brain activity into commands or messages for interactive applications [178], [179]. A typical example of a BCI is a system that enables users to move a cursor on a computer screen towards the left or right, simply by imagining movements of the left and right hand, respectively. Such imagined movements can be recognized from the user's brain activity, typically measured by ElectroEncephaloGraphic (EEG) electrodes placed on the user's scalp. BCIs might find numerous applications, including enabling (severely) motor-impaired users to control assistive technologies, *e.g.*, spellers, power wheelchairs, or prostheses; for real-time mental state monitoring and adaptive interaction (*e.g.*, to adapt the level of automation of an autonomous system to the users' mental workload); or for motor and cognitive rehabilitation, among others [180].

1) Artificial Intelligence in BCI designs and applications

Many BCI systems are based on AI [181], [182] and have been so since some of the first pioneer BCI designs [183]. Indeed, BCIs notably use Machine Learning classification algorithms to recognize the users' EEG patterns associated with a given mental state or intention, *e.g.*, an imagined left-hand movement or a high cognitive workload (a.k.a. mental efforts). Most BCI designs are based on so-called shallow classifiers (*e.g.*, Linear Discriminant Analysis or Support Vector Machines) [181]. Still, like many other fields, a rapidly increasing number of studies and works now explore Deep Learning for EEG classification and BCI designs [182], [184]. However, it should be noted that, contrary to other domains such as image or speech recognition, there has not been, at least not so far, a Deep Learning revolution in BCI. Most international brain signal classification competitions are not won by Deep Learning methods, probably due to a lack of large brain signal BCI databases, but by so-called Riemannian Geometry algorithms [185]–[187], which manipulate EEG signals represented as covariance matrices. This may change in the future, as several efforts are made to gather large EEG data sets by combining the efforts of multiple laboratories [188] or by finding ways to reuse various existing data sets together [189]. Finally, more recently, AI algorithms have been used not only for EEG and brain signals classification but also in the design of the BCI-based application itself, typically to adapt dynamically, in an automatic and intelligent way, the application or interface to the users' mental states. This can, for instance, enable to automatically adapt the level of automation assistance using Partially Observable Markov Decision Process (POMDP) [190], [191]; to propose a personalized sequence of training exercises using Intelligent Tutoring Systems (ITS) [192], [193]; or to learn the users' long-term goals and a model of the users using Reinforcement Learning (RL) [194], [195].

2) Applications of AI-based BCIs

This section presents a brief overview of the various applications of BCI to illustrate its potential and capabilities. It does so by distinguishing invasive BCIs, which measure brain signals from within the skull or even within the brain, using implanted electrodes, from non-invasive BCIs that only use sensors placed on or around the skull [178]. Indeed, these two BCI categories have different capabilities – sensors implanted

in the brain naturally have much better signal quality and information – but they also target different user populations and have more significant risks.

During the last few years, there have been several impressive developments in invasive BCIs, notably those that are based on chronically implanted brain sensors, such as Micro Electrode Arrays (MEA) implanted in the brain (the more invasive sensors) or ElectroCorticoGram (ECoG) implanted below the skull but on top of the brain (thus less invasive). These chronic implants enabled long-term data recording from a given patient, leading to large amounts of data available for AI algorithms and long-term user training, which can also significantly improve BCI performances and capabilities. Indeed, BCI control is a skill that needs to be learned and trained [196]. In the domain of neuroprosthetics, invasive BCIs have notably enabled tetraplegic users to control robotic arms with up to 10 degrees of freedom with MEA [197], [198] or to control a full exoskeleton with up to 8 degrees of freedom with EcoG [199]. Combining invasive BCI with invasive muscle stimulation or brain stimulation to provide artificial sensations also improved the control over neuroprosthetics [200], [201]. Invasive BCIs have also recently made impressive progress in BCI-based communication, with some works enabling an implanted BCI user to spell up to ninety characters per minute, using brain activity only, by imagining the gestures done to write these characters [202]. Even more recently, direct speech decoding was demonstrated, with a spelling speed of 62 words per minute with MEA [203] and 78 words per minute with ECoG [204].

Despite less impressive results due to lower quality signals that are also less informative, non-invasive BCIs, notably those based on EEG, are much more popular and more studied. They are also more likely to be usable by many users since they do not require surgery. The most promising applications of non-invasive BCIs can be gathered in three main categories: assistive technologies [205], Motor and Cognitive Rehabilitation [206], and NeuroAdaptive Technologies (NAT) based on mental state monitoring [207]. Note that there are also works in which non-invasive BCIs can be used for direct control of applications for healthy users, *e.g.*, to control video games [208]. However, the current (un)reliability of non-invasive BCIs (*e.g.*, current EEG-based BCIs that recognize left or right imagined movements make about 20-25% of errors on average, while between 10% to 30% of BCI users cannot use current EEG-based BCIs [209]) make them unlikely to be used in practice outside the lab, at least in the short term.

In terms of assistive technologies, EEG-based BCIs have been used to control wheelchairs (to go forward, turn left or right by imagining, *e.g.*, foot, left-hand or right-hand movements respectively), simple prosthetics (*e.g.*, opening or closing a prosthetics hand), and spellers [205]. The latter is generally based on so-called reactive BCIs [210]. With such reactive BCIs, different letters are displayed on the screen, and a specific visual stimulus is usually overlaid on each letter (a different one per letter). Suppose the user pays attention to one such stimulus. In that case, this evokes a stimulus-specific brain response that can be detected in the user's EEG signals to select the corresponding letter [211]. Latest developments in EEG-based BCI spellers enabled spelling speed at a rate of up to thirty-five error-free characters per minute [212], [213], and have become commercial products, *e.g.*, Intendix from g.tec[8] or MindAffect[9].

FIGURE 5. Illustration of a BCI being used to monitor User eXperience (UX) from EEG signals to assess and adapt 3D interaction tasks accordingly (©Inria / Photo H. Raguet).

Regarding motor and cognitive rehabilitation, non-invasive BCIs are notably promising for post-stroke motor rehabilitation [214]. Indeed, patients who suffer from a stroke may be paralyzed in a limb due to a stroke lesion in the corresponding brain area. Such paralysis may prevent the user from benefiting from physical therapies. However, since BCIs can detect motor intention and imagination from EEG signals, they can be used to guide the patient to use the damaged brain area, even if they are unable to move the corresponding limb: dedicated feedback, *e.g.*, visual feedback of a 3D hand moving on screen or tactile input on the paralyzed hand, can be provided when the BCI detects a movement intention. This can stimulate brain plasticity in the damaged brain area and thus help towards recovery. Recent research and meta-analyses have shown that BCI-based post-stroke motor rehabilitation is clinically effective (in complement to traditional therapy), and it leads to better rehabilitation than motor imagery or functional electrical stimulation therapy alone [214]–[216]. Currently, further applications of BCI-based post-stroke rehabilitation are being investigated, including post-stroke speech rehabilitation [217] or cognitive rehabilitation [206].

Finally, non-invasive BCIs can be used for real-time mental state monitoring (*e.g.*, see Fig. 5). Indeed, these instruments can monitor cognitive, affective, or conative (*i.e.*, related to motivation) states, such as mental workload (related to mental efforts), attention levels, valence or arousal (emotional states), error perception or curiosity [210], [218]–[221]. Naturally, recognizing such mental states is far from perfect, as always with BCIs, with classification accuracies in the range of 60%

[8]https://www.gtec.at/
[9]https://www.mindaffect.nl/

to 90% to distinguish two mental states (*e.g.*, curious vs non-curious or perceived error vs no error). Monitoring such states opens the door to many promising applications, so-called NeuroAdaptive Technologies (NATs), that dynamically adapt to the user's mental state. For instance, NATs can provide optimal sequences of training exercises adapted to the mental efforts of the user for education. They can also be used for neuroadaptive gaming, to adapt the difficulty of the game to the attention or stress of the user, or for intelligent cockpits, to adapt the information provided in the cockpit to the pilot's attention or mental workload, to ensure they do not miss alarm or relevant information [207].

3) Potential benefits and risks of AI-based BCIs

BCIs are very promising for many applications, as they could enable severely motor-impaired users to regain some autonomy and independence, thanks to BCI-based assistive technologies, notably for prosthetics and speller control. They are also promising for the rehabilitation of stroke patients to improve their motor, speech, or cognitive recovery. Finally, mental state monitoring can enhance education efficacy and efficiency, entertainment quality, or safety in critical environments.

Naturally, AI-based BCIs are not without risk either. Since they can monitor brain activity and adapt the interaction with users accordingly, they might be used for ethically debatable applications when not for applications that should be forbidden altogether. For instance, BCIs can be (and are) used for neuromarketing, possibly for tailoring the advertisement shown to users according to their preferences decoded by a BCI, thus influencing –possibly without the users' knowledge- what they will buy. Similar approaches could influence political preferences or, in general, unknowingly manipulate users. To the best of our knowledge, there are no reports in scientific publications that this is feasible in practice, but this seems theoretically a possibility. BCIs could also monitor workers or students (*e.g.*, are they working hard enough based on their estimated mental workload?). Again, these are only possibilities but not realities so far, but these are risks one needs to consider. In general, data privacy is a concern for BCI. Responsibility is another concern: when using an AI-based BCI to control an application if this application (*e.g.*, a wheelchair) creates an accident, who is responsible? These questions and others are ethical and legal questions that would need to be debated and answered [222].

Interestingly enough, the OECD has developed guidelines for responsible innovation in neurotechnologies[10], which include BCIs. Such policies are currently being implemented in the various OECD states. This is a first step towards limiting the risks of AI-based BCI use. However, these are only guidelines which are non-constraining. Actual laws are probably needed, such as the neuro-rights that Chili implemented in its constitution[11]. The recent EU AI act also aims at regulating affective computing, to which some passive BCIs are related [6].

[10]https://www.oecd.org/science/recommendation-on-responsible-innovation-in-neurotechnology.htm.

IV. AI IN SCIENTIFIC RESEARCH

In this section, we examine the state-of-the-art application of AI technology to a selection of scientific research areas, focusing on the crucial elements that are reshaping the field and will drive innovation in the future. In so doing, we also try to identify the most promising and the most dangerous aspects of this impending revolution. Consideration of the latter, in particular, will inform a discussion aimed at imagining pre-emptive measures and possible areas of intervention where we may successfully mitigate or avert the threats posed by a fully unsupervised and unconstrained diffusion of AI.

A. PHYSICS

Physics is traditionally a field of scientific investigation where new ideas and methods emerge and develop only after a difficult gestation period. During that transition, the community goes through a multi-stage process that sometimes may last for a long time. This inertia exists because, in general, physicists are trained to automatically doubt and ruthlessly question new claims and revolutionary ideas. They acquire that attitude by working in a field of science that moves slowly and where genuine revolutions are rare. At the same time, ill-supported or made-up claims of disruptive innovations come

[11]https://en.unesco.org/courier/2022-1/chile-pioneering-protection-neurorights.

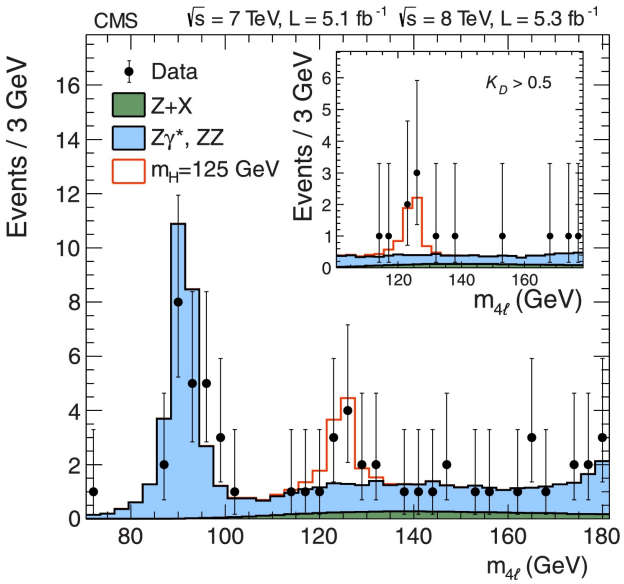

FIGURE 6. The Higgs boson signal (red histogram) overlaid to known processes involving the production of a Z boson and other particles that could mimic the signal characteristics in the "golden" four-lepton final state, extracted by the CMS experiment from 10.31 inverse femtobarns of 7- and 8-TeV proton-proton collisions produced in 2011-2012 by the Large Hadron Collider. Experimental data are shown by black dots with vertical uncertainty bars; blue and green histograms show expected backgrounds.

in at a steady pace. In such an ecosystem, it pays off to err on the side of caution and align with the establishment.

At what we might call "stage zero," a new idea (either a new physics theory, a new instrumental technique, or a new method to extract inference from data) does not receive enough attention to even be worthy of a reaction by the most influential members of the community: it may then remain for a significant amount of time a topic that is only considered and investigated by the few researchers who brought it up in the first place. At this stage, it may be very difficult for the proponents to get the person-power and the resources required to develop their ideas, acquire funding to increase the effort, or publish interim results. Stage one in the path of acceptance takes place when the new idea is finally brought to the attention of a significant number of academics, *e.g.*, by being presented at an international conference or published in visible publications, at a level at which it cannot be ignored any longer. The academic establishment is then likely to express an evaluation or a reaction. The reaction is almost universally one of general skepticism; it may include direct attacks that leverage weaknesses in the less well-developed aspects of the proposed idea, ignoring its strong suits. Proponents at this stage may struggle in the uphill fight to acquire a sufficient consensus from their peers and often get marginalized or even discriminated against[12] [223]. Only when a sufficient part of the community is convinced of the soundness of the new proposed innovation does a transition occur to "stage two," when the idea may be explored with significant resources and start to attract funding and the interest of young researchers. Yet even then, it is quite common for a good part of the community –typically represented by the old-schoolers– to remain critical or even openly opposed to the innovation. Eventually, old schoolers graciously tone down their criticism or simply retire or die. That is stage three, when the new idea and its consequences revolutionize the field.

1) The past

The general multi-stage path outlined above is not dissimilar to the slow-motion rise and adoption that artificial intelligence methods saw in physics research. We may take particle physics as an example to examine such developments. Particle physics, a field marked by a century of intertwined scientific progress and technological advancements, seems ripe for innovation. Despite this, efforts to leverage the rapidly growing computing power that became gradually available at the end of the 20th century faced significant challenges. During this "stage 0", indicatively starting in the 1980s [224]–[226], only a small number of researchers experimented with the new complex algorithms and methods for data analysis that had been made available by recent advancements in computer science and statistics –including statistical learning methods and neural networks; for a review of early studies see [227]. Those pioneering attempts were typically not backed up by the collaborations running large collider experiments. Entering the 21st century, the decade which we may associate with a "stage 1" according to the schematization outlined *supra*, there was a gradual increase in studies of what we now recognize as machine learning techniques, such as decision trees, random forests, and gradient boosting; yet despite this growing interest, the path to publication of analysis results that made use of those techniques remained arduous and uncertain, mainly due to the scepticism of part of the scientific collaborations producing the results, every member of which had the power to request further studies and checks *ad infinitum* before submission to a scientific journal could be made.

Finally, in 2012, a true paradigm shift happened, and "stage 2" started. The ignition point was the discovery of the Higgs boson (see Fig. 6) [228], [229], for which the ATLAS and CMS collaborations at the CERN LHC collider employed machine learning techniques [230]. This coincided with the rise of machine learning in other disciplines, mainly driven by the success of algorithms for image classification in achieving and surpassing human performance. Still, it took a few more years to reach what we might recognize as "stage 3": old-fashioned academics kept questioning for a little more time the indeterminate nature of the mechanisms inside the "black box," the unknown way by which neural networks take their decisions when applied to the classification or regression tasks which are common in particle physics data analysis.

Today, the question, in particle physics as well as in other adjacent fields of fundamental science investigation (including astronomy and astrophysics, nuclear physics, and related physics phenomenology areas), is not any longer the way neural networks –which have grown "deep" even in the most straightforward applications and settings– operate their decisions. Rather, the question now is how to fully exploit the latent potential of the large arsenal of tools developed in computer science and how to customize, extend, and improve the functionality of those new tools for the specific tasks demanded by the investigation of frontier physics. The revolution has fully taken place, and what twenty years ago had moved from being labeled as "statistical learning" (with an emphasis on the properties of the estimators those algorithms could produce) to more properly "machine learning" (with an emphasis on the adjustment of algorithms to the specific learning task, and a focus on their performance), is starting to be called without restraint with the broader umbrella term of artificial intelligence.

2) The present

Artificial intelligence is used today in fundamental physics to improve the performance of experiments, to boost the pre-

[12] A fitting example of this situation is the "dynamical likelihood" technique developed by Kunitaka Kondo and collaborators in the late 1980s. The method was meant to apply to data analysis in the CDF experiment at the Fermilab Tevatron collider, which (among other things) was searching for a signal of the sixth quark, the top. Despite its soundness, the dynamical likelihood method was never accepted for application in CDF analyses, and over a decade had to pass before the technique was accepted as a highly performant, correct method and was finally employed to measure the mass of the top quark. It subsequently became a standard weapon in the toolkit of particle physicists.

cision of measuring instruments, and to achieve sufficiency [13] of summary statistics employed in the analysis of the complex, multi-dimensional data usually acquired by those instruments. In addition, AI is used by phenomenologists to simplify the investigation of large-dimensional parameter space or to improve the potential of theoretical calculations. The gauge of a completed paradigm shift is observing the path of a particle physics analysis to publication, operated by a large collaboration. As we mentioned *supra*, only fifteen years ago a study using a neural network technique would be questioned harshly and would not be allowed to proceed to submission for publication until extensive, excruciatingly detailed, and deep cross-checks were performed. In contrast, today, an analysis result that does not employ a neural network or a similarly powerful supervised learning technique for data reduction and inference extraction will be automatically considered suboptimal, questioned for the reason of a lack of optimization, and often regarded as unworthy of being published.

The above paradigm shift is fueling a sociological effect within the scientific community. Young researchers who start their path as Ph.D. students in physics and dedicate themselves to the extraction of measurements from experimental data are quickly catching the message that they either become experts in those computer science tools or had better move on to some other career path. The result is that physics knowledge is losing value to computer science skills, even within the walls of academia. The long-term outcome of this trend remains to be seen. Still, in the meantime, we can observe how major universities worldwide have promptly reacted to the shift in demand and have started to tap the throughput of bachelor courses in Physics to hire students. The new term "Physics of Data" is being used to advertise or directly name new master courses that focus on computer science concepts and training at least as much as they focus on advanced physics. The message is that to be a good physicist, you must know computer science well today.

Meanwhile, the exploitation of deep neural networks and supervised learning has gradually become only one of the activities on which experiments base the production of their results, as new powerful methods (normalizing flows, transformers, autoencoders) and, in general, unsupervised learning methods have flanked it and gradually grown in importance. Following what is happening elsewhere, deep neural networks are being replaced with the more general concept of differentiable programming, which is the true engine under the hood of neural networks: the capability to backpropagate the gradients of a loss function to optimize the network parameters. Differentiable programming is offering itself for the solution of tasks that, until yesterday, were not even considered as approachable. Here, we mention two of them.

[13]Sufficiency is a property of a statistic (a function of the data) that retains all the information about the parameters of interest. A sufficient statistic provides an efficient summary of the data, often by reducing its dimensionality, without losing information relevant to the estimation or inference of the parameters.

The first novel application concerns generative models. In fundamental physics, the mechanisms responsible for data generation in any relevant instrument are intrinsically stochastic –two protons hitting the same block of matter with the same energy, *e.g.*, will never produce the same chain of reactions. The differences are so stark that the problem is not amenable to solution by writing down a likelihood function of the parameters of interest given the observed data: it is intractable by those means. In such situations, inference can be operated by comparing observation with simulated data that samples the space of latent parameters governing the physics of the system under study; the procedure is captured by the term "simulation-based inference" or "likelihood-free inference." In recent years, new unsupervised learning methods (in particular, normalizing flows, autoencoders, and adversarial networks) have become available to produce generative models that greatly simplify the task of scanning the latent space and extracting precise inference from the stochastic input data. These techniques are thus revolutionizing the analysis of large bodies of stochastic data in fundamental science applications.

The second task AI technology enables is the end-to-end optimization of the instruments conceived by researchers to further their knowledge of fundamental science [231]. Large particle physics or astroparticle experiments involve the design of instruments of such complexity that the only viable option for the study of their potential performance has until now been the discrete sampling of a few "reasonable" configurations informed by experience and the comparison of their performance on a proxy of the actual set of wide-ranging goals of the experiment. This procedure is virtually certain to miss consideration of innovative design solutions that leverage the interplay of any number of the many interrelated construction parameters at play.

The above discrete sampling can be revolutionized by artificial intelligence, exploiting differentiable programming in the construction of a complete model of all the functional blocks of the pipeline connecting the data collection and interaction with the detector with the pattern recognition performed on the resulting detector response and with the successive information extraction procedures, all the way to the calculation of an experiment-wide "objective function" that encapsulates the true goals of the experiment, optionally including an appraisal of constraints, costs, and any other factor worth including in the recipe (see Fig. 7). The differentiable model allows for the complete exploration of high-dimensional parameter space of construction choices in a continuous way, and the identification of the global maximum of the objective function.

There is one additional important point concerning present-day AI's effect on the collateral effects of collaborative research in physics. As in other disciplines, the computing power required for the heaviest applications is quite significant not only in itself but also as a source of carbon dioxide in the atmosphere; it is a source that cannot be neglected any longer when designing and operating large experiments, as

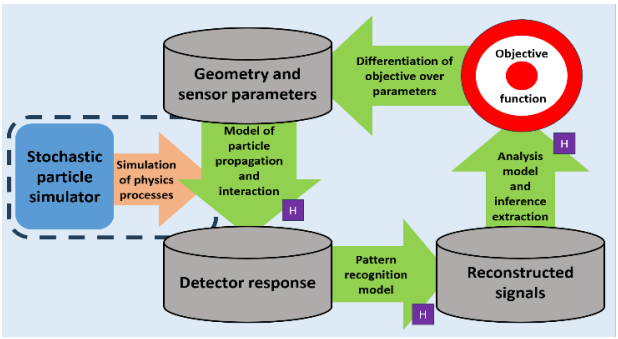

FIGURE 7. Sketch of an end-to-end modeling pipeline encoding all the functional elements that connect the design elements of a detector to the physical processes taking place in it during data taking, the pattern recognition and inference extraction procedures operated on the resulting data flow, and the final computation of a global utility function. Suppose the model is differentiable (*e.g.*, thanks to creating a valid surrogate model of the stochastic processes generating the data, left). In that case, the update of detector parameters (top left) following the gradient of the utility function (top right) allows the instrument to be completely optimized. Boxes labeled "H" indicate where hybrid digital-neuromorphic computing can be integrated into the procedure.

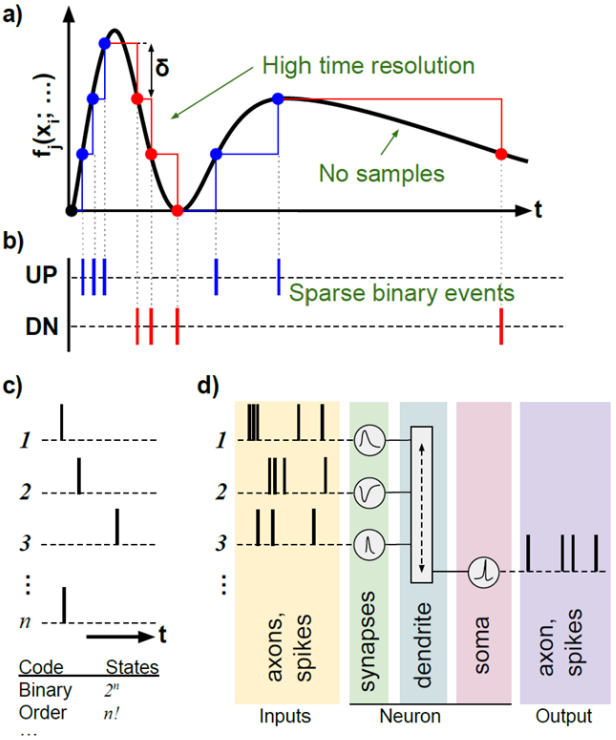

FIGURE 8. Illustration of the encoding and processing of signals by a neuromorphic computing system. In (a), a time-dependent signal (black curve) is processed by Lebesgue sampling to detect the time of integrated changes of its value; this results (b) in two streams of spikes (blue and red dashes) which represent positive and negative variations of the transduced signal. (c) represents the spike-based encoding of the resulting information stream for n channels representing axons; (d) offers a schematic of a spiking neural network unit with input and output channels, mimicking the functional elements of a neural element.

it may exceed by orders of magnitude the emissions due to operation of the experimental complex. For example, analyzing the data produced by high-energy collider experiments requires not only the collection, reconstruction, and storing of collision data but also the concurrent production and use of massive datasets produced by high-fidelity simulations of the physical processes and the apparatus. The ATLAS collaboration recently surpassed the mark of one million 100% active CPU cores employed for their computing effort –a large part of which is used to generate simulated data. As we move to higher-intensity machines (such as HL-LHC, the already approved high-luminosity version of the Large Hadron Collider, which will deliver ten times higher rates of collisions), the impact on our environment of these experimental endeavors is a concern. AI may help significantly reduce this load in several ways. One of them, which is not new, is improving the use of distributed resources by optimizing access to data and computing. Another is the demonstration of the validity of the above-cited generative models, which may eventually restrict to validation tasks the large "full simulation" datasets that today still absorb most of the computing resources of the LHC experiments.

3) The future

Following the two avenues mentioned above to reduce the environmental impact of large physics experiments, a third one that must be placed in the category of prospects is the development of the new paradigm offered by neuromorphic computing. Neuromorphic processors, which encode information with the time of arrival of electrical pulses in close analogy to the behavior of biological neurons (see Fig. 8), have the potential to reduce the energy consumption of computing tasks by many orders of magnitude compared to standard von Neumann computers. For example, it has been calculated that to produce the same amount of computing expressed by a human brain, which consumes about 20 watts of power, a digital computer must employ 20 megawatts [232]. In addition to those vast energy benefits, neuromorphic devices offer co-location of processing (the neuron membrane, which adds potential from the incoming signals and fires once a threshold is reached) and data (the strength of the connections of synapses to the neuron membrane). This constitutes a second disruptive improvement over present-day technology. Neuromorphic computing devices offer a perfect substrate for new artificial intelligence developments due to their natural functioning closely mimicking the biological brain.

Although the technological challenges are significant, no evident showstoppers have been identified in the development of this new paradigm, and the issues appear instead to concern details of its technical implementation. The question, therefore, seems to be not whether neuromorphic processors will be developed and start to compete with conventional digital computers in a significant number of tasks but rather when such a transition will eventually happen. Besides the huge potential of neuromorphic computing in reducing the carbon footprint of large experimental endeavors, its development is also bound to play a quite significant role in many applications for fundamental physics experiments. For example, the

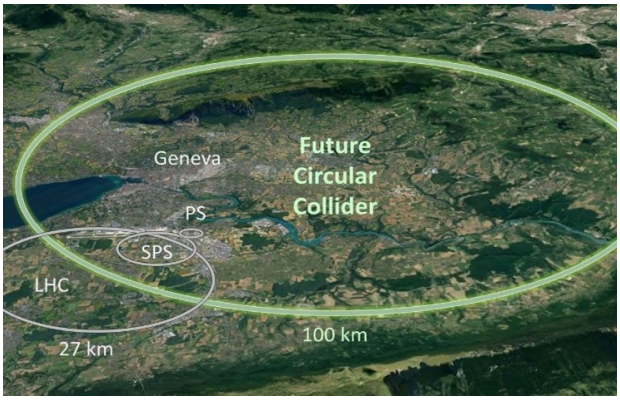

FIGURE 9. **Aerial view of the region around Geneva (Switzerland) with a sketch of the proposed location and dimensions of a future circular collider. The Large Hadron Collider (LHC) and the other smaller facilities at CERN (SPS, PS) are overlaid to the view in the lower left; the future circular collider footprint is green.**

possibility to endow particle detectors with many small independent processors integrated into the detection elements, which perform computing operations with the output of those elements and produce high-level summaries, thus massively reducing the data flow to the backend, constitutes a promising avenue of exploitation of these systems, once their commercial viability and technological readiness reaches the required threshold. Another example concerns applications where the energy needed to operate sensors is a scarce resource (such as in experiments at the north or south pole or experiments in space where a large, distributed set of small, independent units operate detection, pre-processing, and transmission of signals powered by small-scale solar panels): neuromorphic computing is thus poised to provide a new groundbreaking avenue for 21st-century research.

As we point out throughout this work, artificial intelligence is already transitioning to become a more impactful reality in our societies and technology at an increasing pace. This situation, dubbed one of "jolting technologies" (where jolt, a colloquial term for the jerk, is the third derivative of position; a positive, non-zero jerk implies an increase of the force acting on the system, with a divergent effect that we are not accustomed to conceive as it is alien to our physical experience), has been argued to be the signature of an impending singularity [233]. A true singularity would be generated if AI systems were to become capable of autonomously self-improving themselves. This effect is nowhere near being proven at the time of writing. Yet, the impact of this evolving situation is very significant in fundamental physics research because of the large scale and the large time of conception-to-commissioning of the big science experiments required to further our understanding of fundamental physics and of the universe.

For example, a new future circular collider (FCC) has been proposed as the machine that should receive the baton from the Large Hadron Collider after the 2040s, when the HL-LHC phase will have reached its natural termination point. If an FCC or a similar gargantuan-scale machine (see Fig. 9) will be built, it will obviously be entirely based on artificial intelligent systems for the functioning of all its parts (accelerating chain, beam control systems, detector operation, data collection and reduction, event reconstruction, digital twinning and simulation, inference extraction). What is more to the point, though, is that in accordance with the wish to have such a machine ready not much later than the end of operation of the LHC, the community has already started design studies a few years ago. Sitting on a steep slope of technological improvement, however, is not a very comfortable position to be in if we are to make crucial decisions on design, allocation of resources, and other thousands of important choices that play a role in the success of a multi-decennial machine. Therefore, the present situation embodies a misalignment risk. This is not the kind of existential misalignment risk between the goals of humans designing a general artificial intelligence and the goals that the system may decide to set for itself once it is operative (we discuss that situation *infra*, see Sec. V), but it is still a very serious concern. Suppose today we design a machine that will be commissioned and start to take data only twenty or thirty years from now. How can we ensure that we are not making the wrong choices and end up investing efforts and resources in hardware that will be unfit to be best exploited by the artificial intelligence tools that will be available then?

A solution to the above difficulty is to take the "best case scenario" as a baseline –in stark opposition to the attitude one would be advised to take in less long-term projects of this kind. Suppose we cannot model the performance an artificial intelligence may offer thirty years from now in extracting information from generated data by the device we are designing today. In that case, our best bet is to assume that such performance will be arbitrarily large. We are helped by the fact that a ceiling exists for this task: it is the perfect, lossless extraction of sufficient summary statistics from any relevant part of the generated physical processes. In so doing, we are led to concern ourselves with operating design choices that maximize the amount of information physically generated by the system; instead, any choice we made that reduced the information produced by the raw physical processes would be liable to be regretted upon, when using a future higher-intelligence system for inference extraction.

Let us make an example to clarify the point mentioned above. Calorimetric measurements of fluxes of subnuclear particles exploit different mechanisms by which energetic particles lose energy in interacting with dense media. Until recently, calorimeters were only tasked to measure the total energy of incident particles, and their designers did not even attempt to extract from the interactions any usable information on the substructure of the flow of particles crossing the detector, let alone the exact timing of the interactions or useful information on the particles' identity. This construction paradigm was shaken to its roots about twenty years ago, once it was realized how important features could be mined in the substructure of the generated energy flow [230]. Calorimeters

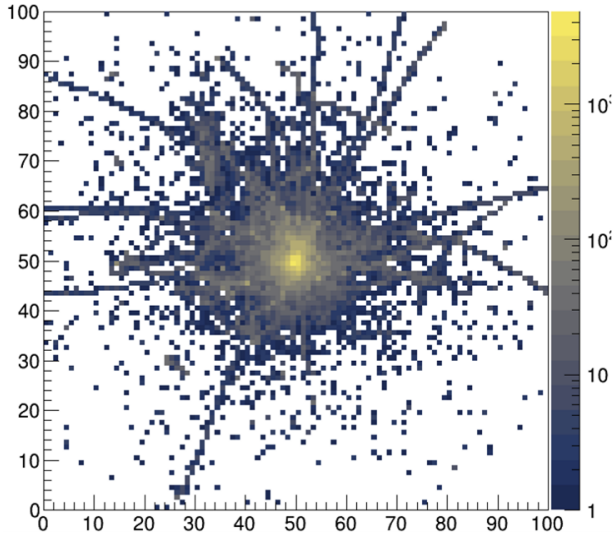

FIGURE 10. GEANT-4 [234] simulation of the energy release in a block constituted by a lattice of 100x100x100 calorimeter cells (of dimensions 3x3x12 mm) made of lead tungstate hit by a 100 GeV proton; displayed units are cell indices. The proton hits the center of the block in the (x,y) projected image shown above by traveling in the z-direction (orthogonal to the xy plane). The lateral diffusion of secondary particles and their energy depositions is visible in the plane orthogonal to the incident proton direction. The released energy (shown by a color temperature map) has been summed over all the cells at the same (x,y) coordinate. It is evident that together with total energy information, the fine-grained cells record information on the number and direction of secondary hadrons produced in the shower, opening the way to the extraction of more information.

suddenly started to be constructed with a finer granularity to gain access to that substructure; experiments that had not considered that aspect at the design stage could not exploit it now. It is already dawning on us that future calorimeters should not only allow access to highly granular information on the longitudinal development of the energy release, as well as its timing (which also produces useful information for discriminating processes of different origins) but they should be built to allow for particles of different identity to manifest their different interaction properties in a detectable manner (see Fig. 10). Such has never been the purpose of a calorimeter in the past, but times are changing fast. While we have not yet been able to demonstrate that it be possible to, *e.g.*, exploit the different localized patterns exhibited by hadrons of different species in interacting with nuclei (*e.g.*, the different interactions undergone by charged pions and kaons hitting nuclei of different materials) for precise particle identification, we must err on the safe side and assume that such a feat will be pulled off in the future, as it is in principle possible.

Whether physicists will adopt the above "forward-thinking" attitude in designing their apparatus remains an interesting question for the coming years. Hopefully, the fast increase in the capabilities of artificially intelligent systems that we are experiencing daily in other areas of human occupation will inspire the present generation of decision-makers in fundamental science and allow them to consider the implication of that accelerated acceleration.

B. FORMAL SCIENCES

The fields of Mathematics and Computer Science are integral to the development and evolution of Artificial Intelligence. This symbiotic relationship enhances our understanding of complex AI models and greatly influences problem-solving approaches in both AI and formal sciences. This section delves into the key intersections between these disciplines, highlighting current research directions, the role of advanced mathematical concepts, and prospects, as they collectively guide the path for AI's theoretical foundations and practical applications.

The theoretical foundations of AI are progressing at an unprecedented rate, as indicated in the survey article [235], which is based on an invited lecture at the International Congress of Mathematicians 2022. The author identifies relevant research directions at the intersection of mathematics and artificial intelligence: mathematical foundations for AI and, conversely, AI for mathematical problems. In this direction, the concept of artificial neurons is crucial for developing deep neural networks. As indicated in Sec. III, Riemann Geometry algorithms are useful in Brain-Computer Interfaces [185].

As an example of the envisioned breakthroughs in mathematics, the use of AI will be crucial to gaining insight into classical partial differential equations and inverse problems. Multi-layer neural networks for deep learning based on fractional differential equations can be used to search an optimal structure [236], and artificial neural networks can efficiently approximate high-dimensional functions in numerical approximations of the Black-Scholes partial differential equations [237].

Let us delve into other innate examples of the relationship between AI and Mathematics and the profound influence that the former is having on the progression of this field of research. The rhythm of this advancement is a well-orchestrated dance of formal sciences and AI, with each complementing and advancing the other. The intersection of AI and formal sciences paves the path for significant advancements in computational complexity, thereby improving the efficiency of learning algorithms and the development of energy-efficient AI. It also buzzes with continual refinement of optimization algorithms, forging the path for enhanced AI learning and problem-solving capabilities. Here are additional realms where Mathematics and AI converge and complement each other.

An understanding of Statistical Learning Theory is crucial for developing machine learning algorithms. Fundamental mathematical concepts like the Vapnik–Chervonenkis (VC) dimension for model complexity, Rademacher complexity to control variance, and empirical risk minimization to reduce training error fortify the application of machine learning algorithms. A recent addition to this blend is Category Theory, focusing on abstract structure and the relationships between mathematical structures. It has found exciting applications in the theoretical foundations of machine learning and AI.

For instance, composing morphisms mirrors the assembly of complex systems from simpler ones, as seen in neural network architectures.

The development and operation of AI algorithms is firmly rooted in the Optimization Method. For instance, techniques like gradient descent, linear programming, and convex and non-convex optimization methods are core to AI algorithms' training process. They play an essential role in managing model accuracy and effectiveness, ensuring an optimal solution is achieved.

Graph Theory forms a pillar for discrete mathematics in AI. Essential for path planning in robotics and strategy in games, it has applications in AI branches like social network analysis and semantic web, specifically ontology. Methods from probability theory and stochastic processes are also vital in AI. Important concepts like Bayesian networks, Markov Decision Processes, and Monte Carlo methods are indispensable for dealing with uncertainty and probabilistic decision-making in AI.

As AI systems increasingly form part of critical operations, their security becomes ever more significant. Advanced cryptographic techniques guarantee secure communication channels and protect sensitive AI system data.

While we still grapple with the fundamentally black-box nature of artificial neurons and deep learning models, Explainable AI (XAI) represents an exciting frontier [238]. It strives to make AI decisions more transparent and understandable. The necessity for an AI agency to oversee and regulate these aspects is becoming starkly evident to ensure the successful coexistence of humans and AI. Developing verifiable and accountable algorithms is paramount to confirming behaviors aligned with desired outputs. This nexus of AI and Mathematics, embracing old and new concepts alike, is key to harnessing AI's potential effectively and responsibly. The growth of this relationship is accelerating faster than ever, indicating an exciting future at their confluence.

Finally, we all are concerned with the black-box-like opacity of AI to detect biases and prevent potential harms [239]. A possibility is to generate an explainable model, a kind of "white box" proxy replicating the inputs and outputs of the original system. An AI Agency to regulate different aspects, as indicated before, is necessary. In all cases, verifiable algorithms should be implemented to monitor and complement the opaque ones.

C. GEOGRAPHY

The application of artificial intelligence techniques to research in Geography dates back to the first geospatial and temporal studies developed by geographers to map terrain more precisely. Already in 1986, Couclelis [240] described how the application of new computing techniques would significantly impact geography. However, it took a few more decades before the power of machine learning brought to the development of entirely new applications [241]. A relevant turning point happened in 2000, when geopositioning through the signal of GPS systems became 10 times more precise overnight, as finally applying a 1996 deliberation of the Clinton administration, the US Air Force removed the scrambling of signals from its satellites. Besides the obvious benefits it brought to navigation systems, the order-of-magnitude improvement in the precision of the localization of devices receiving GPS signals was a powerful enabler of a wealth of new applications. Today, a ground positioning precision of one meter is the standard, and new systems have been announced that will bring the precision down by two orders of magnitude.

A significant further advancement was achieved soon after the geopositioning revolution of 2000, when free availability of high-resolution imagery of the whole Earth's surface was provided to internet users, thanks to a free-distribution software (Google Earth) which was first released by Google in 2001 [242], and has since withstood regular improvements and updates.

Nowadays, the use of geospatial data for mapping, surveys, and other tasks benefits from a number of AI-powered new methods that have been enhancing research. Remote sensing and image analysis are performed with convolutional neural networks, extracting features and patterns from satellite imagery and other sources. This enables improved results in areas ranging from large-scale land cover classification, object detection, and change detection, tasks that may be routinely performed with high accuracy. Geographic Information Systems (GIS) software integrates AI tools to improve data processing, analysis, and visualization capabilities. Besides their use for geography *per se*, GIS applications are also used in various other fields, such as urban planning and transportation management. Environmental monitoring is performed automatically by algorithms that can detect deforestation, monitor wildlife populations, and assess the impact of climate change on ecosystems. In addition, large language models are used to improve data mining from textual and social media data, to understand human behavior and socioeconomic trends at different spatial scales, and to enable studies of the interplay of human activities and the environment.

While the enhancements mentioned above in our capability to harvest and process data from the surface of our planet constitute clear progress and are enablers of better planning and intervention in the environment, there are ethical and social implications arising from the diffuse use of geospatial analysis. Data privacy and algorithmic bias are two issues that apply here, and although they pre-date the application of AI technologies, the use of AI tools is enhancing their relevance.

In the context of artificial intelligence for developing geographical disciplines, a new concept was recently developed: GeoAI [243]. Theoretical advances are now boosted by data management, computer hardware and software, and the fast processing of those data with the new computers. GeoAI is today a discipline within geographical sciences devoted to developing computer programs that can imitate the human view of space and time. Then, the geographical changes, human perception, and spatial and temporal changes are researched using the GeoAI to achieve an advance of knowledge about

the status of the environment and to find solutions to the impact of humans on our planet [244]–[246].

D. AGRICULTIRAL SCIENCES

Agricultural sciences encompass the production and processing of food and fiber, involving technologies related to tillage, plant cultivation, and harvesting, as well as animal production and the processing of plant and animal products for human consumption and use. These sciences face significant challenges, particularly in finding sustainable solutions to the problem of feeding a rapidly growing global population, amidst declining arable land, water, and soil resources due to ongoing environmental degradation and climate change [247]. Agricultural farms and firms must address four main objectives: ensuring an adequate food supply, alleviating poverty, achieving better health and nutrition for a growing population, and conserving natural resources [248]. Historically, agriculture has been a driving force in economic development, playing a central role in agricultural, rural, and structural transformation. Notably, poverty remains most prevalent in rural areas where agriculture provides substantial income, employing 1.23 billion people and supporting over 3.83 billion livelihoods across all stages of the agricultural value chain [249].

Digital crop and livestock farming holds significant potential to meet future food demands [250]. The rapid advancement and diffusion of artificial intelligence (AI) technologies are poised to transform global agriculture. AI, machine learning (ML), and Internet of Things (IoT) sensors that provide real-time data for algorithms are increasing agricultural efficiency, improving crop and animal productivity, and reducing food production costs. Business intelligence research projects that global spending on smart, connected agricultural technologies and systems, including AI and ML, will reach $15.3 billion in revenue by 2025. Farming, traditionally involving numerous manual processes and stages, places immense pressure on farmers. To survive today, farmers must be experts in fertilizers and soils, crop-specific insecticides, planting and irrigation cycles, and weather effects, among other things. AI can complement these applied technologies by facilitating the most complex and routine tasks. It can collect and process large amounts of data on a digital platform, determine the best course of action, and even initiate it when combined with other technologies.

AI in agriculture promises to improve crop management and productivity by phenotyping plants, diagnosing plant diseases, efficiently applying agrochemicals, and providing site-specific agronomic advice. While AI can revolutionize agriculture, farmers need support to implement it correctly. Applying AI in agriculture on a global scale represents a promising opportunity to help farmers minimize or manage their risks to produce economically viable agricultural products.

Bannerjee *et al.* [251] conducted a literature review covering 100 significant contributions from 1983 to 2017, where AI techniques addressed agricultural challenges. In the 1980s and 1990s, rule-based expert systems were widely used in agriculture. As in other fields, machine-learning algorithms have quickly taken the dominant role after the turn of the century. More recently, more advanced hybrid systems such as neuro-fuzzy systems and image processors coupled with artificial neural networks have been utilized.

1) Positive impacts of AI in agriculture

The integration of AI with agriculture offers numerous benefits, including the possibility to quickly and effectively analyze market demand to simplify crop selection and identify profitable agricultural products, the management of risk through forecasting and predictive analytics that may reduce errors in business processes and minimize crop failure, the breeding of crop seeds that are more likely to withstand biotic and abiotic stresses. In addition, autonomous systems may be developed to monitor soil health by diagnosing nutrient deficiencies or toxicity, protecting crops by detecting and managing pests and diseases, feeding crops by optimizing the use of irrigation, nutrients, and agrochemicals, and harvesting crops by automating and predicting the optimal harvest time.

A recent systematic review by Sachithra and Subhashini [252] explored how AI contributes to agricultural sustainability. They found that the most common application of AI in agriculture is predictive modeling for total agricultural output value, followed by harvesting applications. They also noted the increasing use of AI and image-processing techniques to enhance overall efficiency and sustainability in agriculture. In another relevant study, Linaza *et al.* [253] summarized recent research activities through projects that were developed and implemented in a few European countries. They defined terms and concepts related to precision agriculture, described global trends and policies fostering AI-based solutions in the agricultural sector, and reviewed the current state-of-the-art AI applications within precision agriculture. They concluded that AI technologies are instrumental in providing support for decisions for farms, as well as monitoring conditions and optimizing production. For example, they help farmers fine-tune the inputs for each crop. These improvements have positive outcomes in water use reduction, and they may also mitigate the emission of greenhouse gases. The cited source also suggests that substantial improvements may come from developing dedicated, intelligent robots that may autonomously retrieve plant and soil samples or assist in livestock management.

Recent research has focused on improving crop production, preventing crop diseases, managing irrigation, and managing livestock using various AI tools [254]. The integration of high-throughput phenomics and genomics with analytic applications based on large volumes of data has sparked a revolution in agricultural science [255], [256]. AI will enable crop management software to incorporate biological information, developing holistic agronomic programs that integrate both chemical and biological insights. The ability to navigate complexity, distinguish between correlation and causation, and use ML to uncover hidden patterns is crucial for unlocking

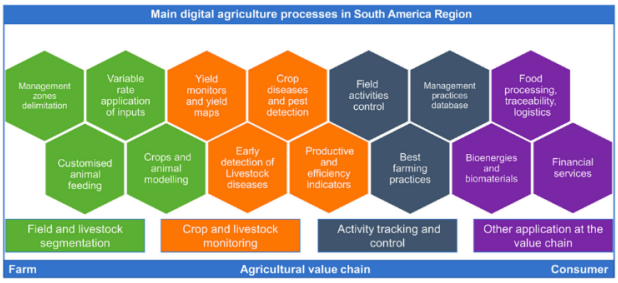

FIGURE 11. Digital agriculture processes within the South America region's farm and agriculture value chain. Processes are grouped into four main categories: field and livestock segmentation, crop and livestock monitoring, activity tracking and control, and other applications within the value chain (Source: [258]).

the big data potential. The simultaneous processing of large datasets provides tremendous opportunities for researchers in academia and industry to advance agricultural science. In this field, the potential of ML models, expert systems, and autonomous machines on farms, farmers, and food security is still largely overlooked or underappreciated.

The large potential of these new methods also entails some risks. Systemic risk factors of AI in agriculture may depend on one side on the interoperability, reliability, and relevance of agricultural data relied upon and on the other on the possible unintended socio-environmental consequences of the narrowness of objectives (such as employing models only optimized for yield), and safety and security concerns with the deployment of ML platforms at scale [257]. Suggested risk mitigation measures include involving rural anthropologists and applied ecologists in the technology design process, applying responsible and human-centered innovation frameworks, establishing data cooperatives to improve data transparency and ownership, and initially deploying agricultural AI in digital sandboxes. Agricultural technology companies (AgTech) are the primary agents of this ongoing revolution toward digital agriculture and the potential drivers of adoption of further innovative AI-based technologies [258]; these companies can provide knowledge-based digital agriculture services at all stages of the agricultural value chain (see Fig. 11). With their expertise and products, AgTech may mitigate the problems and limitations of digital agriculture connected to connectivity, data collection, transmission, storage, accessibility, and interoperability.

Disseminating technical advice and best practices through agricultural extension services is crucial for supporting 570 million small-scale farmers worldwide, contributing to food security and rural development. In a recent study, Tzachor *et al.* [259] assessed the potential benefit to farmers from the use of large language models to transform agricultural extension. This study focused on LLMs' ability to simplify scientific knowledge and provide personalized, location-specific recommendations that may boost productivity. LLMs can be transformative by simplifying science and making advisories personal. Results from the cited study also highlighted the shortcomings of this technology, informed by real-life testing of GPT to generate technical advice for Nigerian cassava farmers. An idealized LLM design process incorporating human experts was proposed to ensure the safe and responsible dissemination of LLM functionality across farming worldwide.

2) The future of agriculture in the AI age

As discussed above, the benefits that AI methods have brought to agriculture are undeniable; AI may integrate diverse types of data across crop and livestock science, unraveling the complex interplay between genetics, nutrition, agronomy, and the environment, as well as influencing on-farm performance and informing ongoing breeding programs for more efficient and sustainable food production [260].

In the future, we anticipate further quantum leaps in agricultural productivity. These may come through the application of cutting-edge technologies such as AI-assisted gene editing, thus meeting today's needs of farmers and the society they feed and anticipating the challenges posed by long-term sustainability in a deteriorating environment and a growing demand. Smart farming tools can perform small, repeatable, and time-consuming tasks, freeing up farm workers for more strategic operations requiring human intelligence. Additionally, the employment of vertical farming systems, which allow the growing of crops in vertically stacked layers in controlled indoor environments, enhances the efficiency and sustainability of crop production. These technological advancements represent the next step in the evolution of smart agriculture, necessitating other technologies to function effectively. Technology providers must address challenges related to improving their tools, helping farmers solve them, and communicating how ML helps solve real problems, such as reducing manual labor. As farmers, cooperatives, and agricultural development companies embrace data-centric approaches and expand AI use to improve yields and quality, the future of AI in agriculture looks promising. Wider adoption of AI-based agricultural practices will only be achieved through collaboration among researchers, technology developers, suppliers, farmers, and their advisors across the digital innovation system [261]. AI will unlikely replace the farmer, but it will significantly enhance decision-making, transforming agriculture into a more efficient, sustainable, and climate-resilient future.

E. MEDICAL SCIENCES

1) The past

The history of AI in medicine dates back to the mid-20th century and has gone through various phases of development since. AI's popularity rose in 1950 when Alan Turing introduced the idea of computers mimicking human intelligence in his book, "Computer and Intelligence" [262]. In the 1960s, Stanford University developed DENDRAL –an expert system that could analyze organic chemical compounds by examining their mass spectra and applying general chemistry knowledge. This was one of the first AI expert systems to be integrated into medical applications as it replicated a chemist's problem-solving and decision-making processes

[263]. Also, during this time, the SRI Artificial Intelligence Center created Shakey the Robot in 1966, which could plan its own routes and rearrange standard objects using humanistic thought processes such as perception and reasoning. Referred to as the "first electronic person" by Life magazine, Shakey highlighted the soaring potential for AI to understand and execute instructions [264]. A decade later, Stanford's School of Medicine developed MYCIN, a tool assisting physicians in determining the bacterial pathogens behind infections such as bacteremia or meningitis, as well as recommending the appropriate antibiotic therapy dosage options according to the patient's mass [265].

Indeed, the 1970s were a pivotal decade for AI in medicine. Already in 1971, the University of Pittsburgh's School of Medicine developed INTERNIST-1 [266], a large-scale set of disease profiles containing patient symptoms, laboratory abnormalities, signs, and demographic data (Institute of Medicine [US] Council on Health Care Technology 1988). This database was designed to assist medical professionals in diagnosing complex diseases. In 1976, the LDS Hospital in Salt Lake City introduced the Health Evaluation through Logical Processing (HELP) system [267]. HELP was the pioneering hospital information system that combined the collection of clinical data with computerized aid in real-time decision-making[14]. This provided medical professionals with alerts for harmful drugs and other clinical recommendations. Despite being innovative and showcasing the potential for AI in medicine, these systems struggled to gain widespread acceptance due to certain technological constraints and skepticism from many groups within the healthcare industry and the public.

The 1980s and 1990s saw continued progress with significant advancements. DXplain, a system developed by the Laboratory of Computer Science at the Massachusetts General Hospital in 1986, could provide justifications for over 2600 diseases based on clinical symptoms and user laboratory data. Specifically, DXplain can output a ranked list of potential diagnoses to be considered based on symptoms, recommends which additional information should be collected, as well as manifestations that would be atypical for each specific diagnosis[15]. AI applications extended into cardiology with consultation systems and clinical tools like CorSage, developed by Cedars-Sinai in 1989 [268]. CorSage utilizes a blend of AI and statistical methods to assist physicians in pinpointing heart patients who are at the highest risk of experiencing recurrent coronary events. Further, The Human Genome Project, launched in 1990 by the National Human Genome Research Institute, provided crucial genetic data that allowed AI systems to understand further genetic factors in diseases[16]. Numerous organizations, projects, and hospitals played a crucial role in supplying data to train and operate these AI systems.

[14]https://www.clinfowiki.org/wiki/index.php/Health_Evaluation_through_Logical_Programming_(HELP).

[15]https://www.mghlcs.org/projects/dxplain.

[16]https://www.genome.gov/human-genome-project.

Machine learning caused a significant evolution at the dawn of the 21st century when neural networks revolutionized medical diagnostics, particularly the use of CNNs in imaging. For example, regarding breast cancer detection, Google's DeepMind developed a CNN model that understands mammograms to identify cancerous lesions with an accuracy that often surpasses humans[17]. Another relevant application is in diagnosing diabetic retinopathy. In 2020, Temple University developed IDx-DR, the first FDA-approved autonomous AI system for this purpose, and it uses CNNs to analyze retinal images and accurately detect diabetic retinopathy [269]. Overall, this evolution of AI highlights its transformative influence on medical diagnostics and its growing acceptance within the healthcare industry.

2) The present

Today's applications of AI in medical science can be categorized into two main branches: virtual and physical. We examine them separately below.

The virtual part includes a wide spectrum from electronic health record systems to neural network-based advice providers in clinical decision-making. AI-driven electronic health record systems incorporate features such as Natural Language Processing (NLP), intelligent image input suggestions, and data entry recommendations to improve operations efficiency and simplify patient record management for healthcare professionals [270]. Another virtual aspect is rooted in Machine Learning, which harnesses algorithms and data to enable AI systems to replicate human learning processes, enhancing their precision over time. Recent developments in AI have improved the prediction, speed, efficiency, and accuracy of the diagnostic process. AI algorithms can analyze large amounts of medical data such as vital signs, biosignals, medical history, laboratory test results, demographic information, and imaging data such as MRIs, ultrasounds, X-rays, DXAs, and CT scans. This analysis helps medical professionals make informed decisions and accurate predictions in patient care [271]. An example is IBM's "Watson for Oncology" software, which oncologists use to make treatment decisions for cancer patients [272]. By combining attributes from data from Memorial Sloan Kettering and a patient's file, Watson for Oncology recognizes and orders possible treatment plans. Other medical disciplines, such as drug development, digital consultation, and pathology, use AI to assist medical professionals.

Machine Learning has facilitated advances in medical sciences. For example, unsupervised algorithms for protein-protein interactions contributed to the discovery of promising therapeutic targets [273], computational methodologies for recognizing DNA variants (*e.g.*, single nucleotide polymorphism) to predict specific diseases or physiological traits [274], and particular algorithms for electronic medical records to capture and process data in real-time to facilitate the finding of patients with a positive family history for spe-

[17]https://health.google/caregivers/mammography/.

cific genetic disorder or individuals with increased risk of specific chronic diseases [18]. Recently, ML models have proven beneficial in assisting healthcare professionals in diagnosing various diseases and illnesses earlier based on a series of parameters. Another interesting virtual application of AI in medical care is the utilization of softbots. Softbots have been introduced as teachable psychotherapeutic avatars and have shown promise for pain control measures in pediatric patients with cancer and for detecting emotional disturbances such as suicidal ideas [275].

The physical part includes advanced medical devices, intelligent prostheses, and complex robots for care delivery (carebots). AI-assisted robots have been used as companions for the geriatric population with cognitive or mobility impairments, as assistants in surgeries or solo performers, and as teachers for autistic children [276], [277]. Nevertheless, routine application of AI-assisted robots is associated with major ethical issues and requires standardization, precise evaluation of the efficacy, and close follow-up on the related side effects and outcomes.

The evaluation of patients' performance in rehabilitation is improved by using AI systems such as wearable health sensors. Interestingly, AI is suggested as a powerful tool for monitoring guided drug delivery to different tissues.

Some AI-based tools received FDA approval; for example, Kardia for ECG monitoring on smartphones and detecting atrial fibrillation [278], Guardian and Sugar.IQ systems for glucose level monitoring and prediction of hypoglycemic episodes [279], [280], Empatica's wearable Embrace for detecting epileptic seizures [281], [282], wearable sensors for evaluating gait, posture, and tremor in patients with neurodegenerative diseases [283], and other algorithms for interpreting pulmonary function tests [284], [285], predicting the decrease in glomerular filtration rate [286], processing the endoscopic and ultrasound imaging in gastroenterology [287], and assisting cancer diagnosing via computational histopathology [288].

3) The future

Using past data to improve clinical decision-making is the core concept of evidence-based medicine. AI facilitated uncovering complex associations among a vast set of different data, which cannot be humans. This allows machine learning systems to approach complex clinical problems in a way similar to how a physician would operate. The superiority of these systems to clinicians is the openness to numerous simultaneous inputs and the ability to rapidly process data. Indeed, such systems can obtain data from as many cases as a clinician may visit in a lifetime, just in a minute. Accordingly, AI-assisted tools were successfully used in the classification of suspicious skin lesions, radiologically determining pulmonary tuberculosis, and triaging systems [289]. Taken together, AI performance in well-defined tasks, such as the classification of suspicious skin lesions, had higher sensitivity and specificity than expert dermatologists. Moreover, AI could be a promising tool for poorly resourced services; for example, the lack of expert radiologists in remote regions where tuberculosis is prevalent could be compensated by deep learning models [290].

The use of AI tools may reduce the financial burden on the healthcare system while also offering more personalized clinical advice to patients. Application of AI-assisted tools in medical settings may reduce the workload of healthcare staff, hence they can spend more time on critical cases.

AI systems do not have empathy and compassion; hence most patients prefer to be visited by a human physician. Lack of direct human-human contact may result in imperfect communication and lead to the system missing essential information that human physicians may be able to observe in addition to the main patient's complaints and details of their health problems; that ancillary information may sometimes prove necessary for clinical decision-making. In general, AI systems are still less trusted by patients [291]. On the other hand, uncontrolled access to AI-assisted medical help may result in misuse, information overload for the patient, misdiagnosis, and mismanagement.

Some serious criticism focuses on the validation process of AI tools. For instance, most of the studies on the efficiency of AI versus human physicians are suggested to have unreliable designs and lack primary replication [292]. In addition, the majority of the studies on the use of AI in clinical practice are criticized due to retrospective design, sample size, spectrum bias, and selection bias [287]. Finally, confronting AI and physicians is probably not the proper approach; some studies considered the combination of AI-assisted tools and human physicians, which outperforms either alone [293].

Different ethical issues are raised with AI development. Considering that medical sciences are engaged with people's health, addressing ethical issues concerning the application of AI in medicine is of utmost importance. Again, we insist that an AI Agency to supervise these ethical aspects is necessary.

AI and data mining unlock vast possibilities for uncovering new links between health and various aspects of life, drawing from extensive repositories of health records (*i.e.*, US Medicare, US Veterans Affairs, British Tissue Repository, ADNI, and others) and genomic data available on scientific and genealogy platforms such as 23andMe, Ancestry, and similar sites. Despite official governmental restrictions on access and use, there are data privacy concerns associated with many of these repositories. In December of 2022, BlackRock purchased Ancestry.com[19] and later a couple of plaintiffs filed a class action complaint, alleging BlackRock violated the Genetic Information Privacy Act[20]—a California legislation that imposes privacy regulations on direct-to-consumer genetic testing companies while also affording consumers access and

[18]https://doi.org/10.1186/s13336-015-0019-3.

[19]https://cookcountyrecord.com/stories/641978335-appeals-panel-blackrock-s-purchase-of-ancestry-com-doesn-t-mean-they-can-be-sued-for-obtaining-illinoisans-genetic-info.

[20]https://privacyrights.org/resources/genetic-information-privacy-act-california.

deletion rights. The plaintiffs alleged that before the Blackstone acquisition, Ancestry linked their saliva sample genetic sequencing kits with personal details such as email and home addresses. They claimed that Blackstone's acquisition compelled Ancestry to disclose this protected information. Despite official restrictions on access and use, concerns arise when investors like BlackRock heavily invest in platforms such as Ancestry, prompting questions about the potential unauthorized transfer of data to insurance databases.

Independent from investing in repositories, further data privacy concerns exist, such as the logistics behind training AI models. High-quality, large, and diverse data sets are required to effectively improve and train AI models; however, not all organizations collecting medical data are shielded by the HIPAA privacy law. Furthermore, erroneous or biased data used by AI can lead to inaccurate predictions, causing under or over-diagnosis and treatment[21]. The question remains: when will the precision of machine learning exceed that of the top specialist expert? Diagnosis is the gateway of treatment plans, and here, AI possesses the capability to process exponentially more inputs than the human brain, facilitating world-class diagnosis and treatment to be delivered anywhere while highlighting the lack of infrastructure required for delivery.

In Alzheimer's disease (AD), for 30 years, researchers have evaluated the most highly predictive genetic factor in AD, the apolipoprotein genotype (APOE). APOE is a reliable predictor of Alzheimer's disease risk, with individuals carrying the E4 allele facing three times the risk and homozygotes facing a risk 10 to 20 times greater. Add this to the differential penetration among African Americans and Hispanics, and there are clear risk classes [294]. The use of AI to mine large demographic and patient-record databases can reveal additional risk predictions. These analyses extend beyond diagnostics, accurately predicting individuals who are not yet affected by the condition. This raises ethical concerns surrounding issues related to information accessibility, counseling restructuring, and the stigma associated with pre-symptomatic labeling. These matters necessitate vigorous dialogue in this rapidly emerging area, which not only provides labeling from databases and genetic data without individual knowledge but also from speech or gait patterns. Developing policies for most conditions requires independence and should not solely depend on patient advocacy groups, which agents of industry frequently back.

As AI tools are trained based on specific datasets, the algorithms may be biased toward certain patient groups, leading to discriminatory traits. In addition, AI tools need access to a vast amount of patient data to be trained and to make decisions. Hence, data security and privacy, as well as data ownership using AI tools, is controversial. Along the same lines, patients probably do not understand how their data is being processed by AI tools, which endangers informed consent. There are also concerns regarding the accountability (the responsible party for AI decisions and actions in the healthcare setting) and transparency (the complexity of AI systems makes it challenging to rationalize how the system arrived at a decision) of AI tools. From the global health perspective, the application of AI-assisted tools may exacerbate the current health disparities. Another challenge would be the overreliance on AI systems, which may lead to debilitating critical thinking and decision-making skills of healthcare professionals [295], [296].

AI can transform the grounds of medicine as it is time-efficient, inexpensive, and does not have physical limitations as humans do. It offers great promise to improve healthcare and reveal the root causes of genetic, lifestyle, and chronic diseases. However, parameters must be established to address patient data privacy and ethical concerns. Medical professionals must provide AI control, oversight, and security to utilize its transformative potential in medicine fully.

F. BEHAVIOURAL SCIENCE

Behavioral science offers an example of how AI has the potential to help scientific progress as well as to hurt and hinder it. We discuss this contrasting situation below.

1) Empowerment of behavioral science by AI

Artificial Intelligence has the potential to significantly augment scholarly work in behavioral science, offering tools and methodologies that can enhance the efficiency and effectiveness of research and practice.

Among the more obvious benefits is the potential efficiency of uncovering past knowledge if the models are properly curated. Advanced search algorithms can sift through vast databases of scholarly articles, rapidly extracting findings that are relevant to specific questions. With enough curation, the quality of research can be weighed as part of this process.

Barriers to fair consideration of research findings can potentially be reduced in this fashion. For example, standard indexing engines have a known bias against non-English sources or studies from lower-and-middle-income countries –even if the science is of high quality [297]. Biases of this kind can considerably distort systematic reviews of the literature [298], but AI tools can help overcome these language barriers.

Some research topics that would simply be impossible without AI can be explored. For example, communications researchers might study the impact of discussing modern issues in the style of historical figures (*e.g.*, "What would Walter Cronkite say about abortion and how would it land?"). Other research topics might be explored with greater experimental control. For example, social persuasion research could examine the impact of AI-generated transcripts in which scientists tightly specify the parameters. The ease of conducting experiments could also be increased by using AI tools to manage and analyze large datasets, automate recruitment, or manage the engagement of participants.

Access to sophisticated modeling tools is another advantage afforded by AI that is already widely in use. More

[21]https://veterans.house.gov/calendar/eventsingle.aspx?EventID=6371.

advanced statistical and measurement tools can better detect complex behavioral patterns, allowing for the development of more precise theoretical models. For example, extensive high-density longitudinal data on scores of individuals with specific mental or behavioral problems can readily overwhelm conventional models, but AI methods can discover signals that reside amidst a great deal of noise. AI tools applied to longitudinal data on the process of change and their relation to outcomes in a large number of individuals, each considered ideographically, may be able over time to produce a more functional diagnostic system than the current normative categorical systems such as the Diagnostic and Statistical Manual of the American Psychiatric Association [299]. As this occurs, kernels of specific evidence-based interventions might be suggested or delivered in a just-in-time fashion that better meets the needs of the individual.

2) Empowerment of contextual behavioral science research

We will now provide a concrete example of how behavioral science research can be empowered by AI by using the Contextual Behavioral Science (CBS) research program as an example. CBS embodies a distinct approach within the broader behavioral sciences that emphasizes understanding behavior in relation to its historical and situational contexts [300]. This approach is multilevel, process-based, multidimensional, prosocial, and pragmatic. The core characteristics of CBS can be described as follows:

- Multilevel: CBS research examines behavior across different levels of analysis (*e.g.*, biological, psychological, social) to understand how these levels interact and influence one another. It aims to address the coherence across these levels within a broad evolutionary science framework.
- Process-based: CBS emphasizes identifying and understanding processes of change that are functionally important. This involves focusing on dynamic sequences of events that can lead to clinically relevant outcomes, linking these to basic behavioral and evolutionary principles.
- Multidimensional: CBS research approaches human behavior from various dimensions, including cognitive, affective, behavioral, and social aspects. It considers the complex and interrelated nature of these dimensions in understanding and influencing behavior.
- Prosocial: CBS research is guided by a prosocial purpose, seeking to foster social justice, cooperation, and the betterment of society. It involves being explicitly aware of the societal impacts of research and aiming to address issues like equity, fairness, and inclusiveness.
- Pragmatic: The approach is inherently practical, focusing on developing strategies and interventions that directly and meaningfully impact real-world problems. CBS values research that can be applied to improve human well-being and adapt to changing societal needs.

In essence, CBS research is about understanding behavior in context, emphasizing practical application and social relevance, and striving to create an adequate behavioral science for addressing the complexities of the human condition.

As mentioned in Sec. III, the Non-Axiomatic Reasoning System and similar AGI systems offer a unique opportunity to enhance our understanding of human cognition and behavior from a CBS perspective. Given that NARS learns in interaction with its environment (rather than being pre-trained on data sets), it could, in principle, be used as an experimental "subject" in settings and procedures akin to those used with human participants. This research strategy has been used to study the development of various forms of concept formation and relational reasoning abilities [301]–[303]. This presents a novel approach to studying multilevel interactions and processes. NARS, designed to mimic certain aspects of human reasoning, provides a controlled environment where researchers can manipulate variables and observe outcomes in a way that is not possible with human subjects. Hence, NARS can serve as a dynamic model for understanding how different levels of analysis –such as cognitive processes, whole-organism behaviors, and social interactions– interact with each other to produce complex behavior patterns. By embedding NARS in simulations that mimic real-world social settings or psychological conditions, researchers can manipulate and observe the impact of changes at one level (*e.g.*, altering cognitive rules or input stimuli) on other levels (*e.g.*, behavioral responses and social dynamics). This approach allows for detailed experimentation on complex, multilevel phenomena in an ethically permissible and highly controllable way. The insights gained from such studies could then inform the development of hypotheses about human behavior, which can be tested in naturalistic settings.

Furthermore, in principle, NARS could be used to carry out process-based research regarding clinically relevant change processes. NARS is unique in that it has a concept of "self" [304]. The SELF in NARS embodies a critical component of its artificial general intelligence, enabling adaptive behavior and decision-making in complex environments. This concept facilitates self-awareness and self-control within NARS, allowing it to perceive and interact with its internal environment similarly to its external surroundings. The SELF-concept evolves through the system's experiences and interactions, starting with built-in operations and expanding through learned behaviors and modifications based on feedback. As a dynamic and evolving feature, it enriches NARS's functionality, enhancing its autonomy and its ability to refine operations and behaviors to better meet its objectives and respond to environmental challenges.

The application of NARS in process-based research offers a novel approach for investigating the cognitive and emotional dynamics involved in, for example, anxiety disorders. This methodology provides experience to NARS that leads to it developing a self-concept that encapsulates various anxiety-related cognitive processes and emotional states, such as worry, avoidance behaviors, and safety-seeking actions. By arranging a series of anxiety-inducing procedures, re-

searchers could manipulate and measure changes in NARS' internal state and behaviors in response to interventions like cognitive-behavioral therapy (CBT) techniques, exposure therapy, and mindfulness-based interventions. The use of NARS would allow for detailed tracking and analysis of how interventions impact anxiety levels and behavioral responses over time, with a particular focus on how changes in self-awareness and the effectiveness of previous coping strategies influence future behavior and self-concept adjustments. This process-based approach would facilitate a deeper understanding of the mechanisms underlying anxiety and its treatment and provide insights into the generalizability of coping mechanisms across different scenarios. Such findings could significantly enhance the development of therapeutic techniques, emphasizing the role of self-awareness and adaptive changes in self-concept in anxiety management. Refer to [304] for an example regarding fear learning.

In addition, AI, in general, can advance the prosocial aims of CBS research. By leveraging AI-driven analysis, researchers can better understand the societal impacts of their work and identify strategies that promote equity, fairness, and inclusiveness. AI can help in modeling the societal implications of various interventions, thereby guiding the development of more socially responsible and effective approaches.

Lastly, the pragmatic nature of CBS research is well-served by AI's ability to provide practical, data-driven insights. AI can aid in rapidly prototyping and testing interventions, streamlining the process of translating research findings into real-world applications. This can significantly enhance the ability of CBS researchers to develop strategies that directly address human well-being and adapt to societal changes.

In conclusion, AI offers a suite of tools and methodologies that can significantly enhance the multilevel, process-based, multidimensional, prosocial, and pragmatic aspects of CBS research. As such, NARS, as it applies to CBS research, is an example of an AI framework that could potentially enhance how behavioral science research is more generally conducted. By integrating AI into their toolkit, researchers can achieve a deeper, more comprehensive understanding of behavior and develop more effective interventions to improve human well-being within a broad evolutionary and life science framework.

3) Risks to behavioral science from AI

Despite all its benefits, AI also introduces risks that could undermine the integrity and progress of behavioral science. One primary risk is the propagation of false information that can be mixed with factual data in ways that are difficult to detect. The ease with which AI can generate content further poses a significant threat to the originality and credibility of scholarly work. Plagiarism is a constant threat.

Depending on how models are developed, the risk of cultural hegemony and bias is embedded within AI algorithms. Commercial AI tools could easily be curated in ways that favor specific commercial products or areas, such as research tools that ignore design problems in research on pharmaceuticals for behavioral and emotional problems while criticizing them in psychosocial research on these topics.

There is a risk that the outcome success of large language models could be mistaken for an understanding of the natural processes of language acquisition, especially in vulnerable populations such as children with disabilities. The "Chinese room" problem in this article shows the issue. Still, confusion of this sort could detract from the search for the human processes underlying language competence, thus undermining effective application and slowing knowledge development.

In the area of psychotherapy, as LLMs come to do a better and better job of mimicking the actions of therapists, the human therapeutic alliance may come to be replaced by machine $\rightarrow$ human substitutes rather than to use AI tools to extend and augment therapeutic interactions. Famous psychotherapists may have their voices and mannerisms, in essence, taken away from them as part of this change. The long-term impact of AI therapists is hard to model in the absence of data, just as with the effects of humanoid robot substitutes in other areas, but it raises difficult ethical and moral issues at the very least.

Finally, the allure and ease of use of new AI methodologies may overshadow the foundational purposes of research and the hard work needed to produce new scientific knowledge. For example, a young scientist might find it far easier to conduct AI-driven meta-analyses of other scientists' work than to risk failure in discovery-oriented research in the lab. This could slow scientific progress even if AI tools perform very well because they cannot replace the need to uncover the processes that lead to behavioral competencies through actual experimentation.

4) Possible mitigation strategies

Several strategies can be implemented to minimize the risks associated with AI in behavioral science. The creation of sophisticated tools to detect plagiarism and cultural or other biases within AI-generated content is essential. In a similar way, improving the capacity to discern false information through enhanced AI algorithms may help improve scientific quality.

Psychiatry has long pursued better conflict of interest (COI) regulations to manage the influence of commercial interests on research outcomes, but these can become "pro forma". In AI, merely waiving one's hand will not suffice. The involvement of scientists and professional associations in setting standards and providing guidelines for AI use is critical to ensure that the integration of AI into behavioral science enhances rather than detracts from the field's integrity and progress.

V. IMPACT OF AI DIFFUSION IN TODAY'S AND TOMORROW'S SOCIETY

A. ETHICAL ASPECTS

The introduction of AI in today's human society raises concerns that are even more radical than those championed in

the XIX century by the Luddite movement, which strongly opposed the introduction of machines in the textile industry.

While most critics of the introduction of machinery in XIX century industries used to focus on the damage to the overall employability of the workers that were being made redundant, we know that the work market somehow adjusted itself to a new configuration. A side effect of utmost importance, however, became apparent when industrial production evolved to Henry Ford's concept of supply chain: the alienation of the workers, whose job changed from a creative activity where "a cycle" was the creation of a whole finished product, to a repetitive fast activity where the job collapses to a single, quick, repetitive action that requires no thinking nor any critical or creative thought.

The changes to society were radical: the frustration of the workers and their exploitation created the conditions for the rise of new ideologies and self-consciousness of the workers [305]. The rise of Communism molded most of modern politics and history: like all human activities, it involved lights and shadows: most of the laws that regulate the job market nowadays (at least in places that did not adopt the ultra-capitalistic system of the USA) came as a result of the activities of the communism-inspired Unions; the implementation of the Communist ideology has however also brought perversion, mass murder, and inequality in many countries.

The AI revolution can have an even more massive impact than the industrial revolution. AI promises to make most of our actions faster or simpler and is often used to improve connectivity between humans. Still, each benefit is unfortunately open to being exploited in perverse ways. The widespread terror about the impact of AI on the job market has inspired several populist movements across Europe to make "citizen income," a guaranteed income that each citizen would receive to be able to live a normal life even without working, a central message of their campaign (cite Italy, Spain, France, etc.). When applied to human interactions, AI can increase connectivity but also be used to manipulate people's minds and political orientation, enabling companies or malicious actors to sway and decide the result of elections, with negative impacts on the population, increasing overall unhappiness. Several tasks that were once exclusively performed by humans can be executed quickly by large language models today, but this entails several risks. Although non-experts use them to produce content, interpreting their output requires expert knowledge to spot errors correctly. This significantly impacts the reliability of AI-produced content, particularly when humans do not honestly declare the AI generation aspect. AI content and the spreading of its use can be devastating from the pedagogical point of view even when the content itself is correct and reliable: students may be tempted to use the outputs as a substitute for putting real effort into studying a topic, and the knowledge of the subject matter will be more and more delegated to the algorithm, to the point that maintaining a certain level of human know-how and control may become unfeasible.

Among the various essays, Yoshua Bengio *et al.* [306] stands out as one of the most authoritative. Its contributors include prominent intellectuals such as Yuval Noah Harari and Daniel Kahneman, alongside key figures in the current AI revolution, including Geoffrey Hinton. Despite their significant contributions to the advancement of AI, these authors express both surprise at the rapid progress and an anticipation of new frontiers that AI will soon reach. They argue that the unpredictable extent of AI's potential intelligence shortly necessitates a severe examination of its associated risks. Rather than calling for a slowdown in AI research, these authors advocate for an accelerated focus on integrating safety measures and verification processes into AI development.

1) Benefits and risks of AI diffusion to modern society

AI-generated content and advice are widely used in sectors like healthcare, education, literature, report production, language translation, and social media [307]–[312]. One of the major benefits of AI is that it significantly simplifies content production, both written and visual. Advances in machine learning and AI have led to language models (see *supra*, Sec. III) that generate credible continuations to short prompts. For example, these applications are revolutionizing journalism. They automate writing tasks such as local news stories, special interest stories, and earnings reports. This allows human efforts to focus on editing, selecting, organizing, and presenting content [313].

The rise of AI has also sparked concerns over the proliferation of fake news and misinformation. AI can be used to identify those susceptible to misinformation and manipulation and to create content that sounds reasonable and mimics human news stories, narratives, and behavior [313], [314]. For example, social bots significantly contribute to the spread of low-credibility articles. They amplify such content in the early stages of spreading before it goes viral. These bots target users with numerous followers by responding to them and mentioning them [315]. The users, in turn, reshare the content posted by bots, causing fake news to go viral.

Fake news refers to fabricated information that imitates legitimate news media in form but lacks the organizational process or intent [316]. Unlike legitimate news outlets, fake news sources do not adhere to editorial standards that ensure accuracy. Misinformation can also come from conspiracy theories and health or vaccination misinformation [317], [318]. Studies show that misinformation spreads more widely and quickly than accurate information [319].

During the US presidential election, a quarter of tweets were either fake news (10%) or extremely biased (15%) [320]. Automated accounts and bots played a significant role in spreading fake news. A survey found that 36% of respondents believed in conspiracy theories about the planning of the coronavirus outbreak by powerful people [321]. Research suggests that a substantial proportion of the public views misinformation as highly reliable [313].

Right-leaning media was more linked to fake news than left-leaning media, potentially influencing the election and democratic processes [320]. Misinformation can also lead to

false perceptions, risky behavior, and distrust of authorized information [322]. Health risks can arise from misinformation about such things as vaccines, the Zika virus, Lyme disease, and Ebola prevention and treatment strategies [317].

In summary, while AI can aid useful content, it can be used to support the generation and spread of harmful misinformation.

2) Mitigating risks of fake and misleading information

Fake news and misleading advice are intentionally written to mislead readers with false information, making it challenging to detect based on content alone [323]. Therefore, it is crucial to consider auxiliary information to identify fake or misleading information. This can include user likes, social media stances, and user content annotations.

There are three major ways to fact-check information. First, information can be checked by experts –a resource-intensive task. Second, crowdsourcing can be used to enable regular users to annotate content. These annotations can be aggregated to produce an overall assessment of the accuracy of the information [323]. This approach depends on crowds being "wise." Finally, computational and AI-oriented information-checking can identify unreliable, misleading, or false information [324]–[327].

The use of AI in identifying false information establishes a delicate balance between technology and privacy, data protection, transparency, and explainability of the AI system [328], [329]. AI's effectiveness in detecting disinformation relies on analyzing large data sets, often involving intrusive access to users' online activities and social media interactions. This raises significant privacy concerns, including potential data breaches and the risk of false positives that could stifle free speech. Moreover, who controls the AI and the criteria for identifying "fake news" adds another ethical layer; a clear pitfall to be avoided is the involvement of private corporations in the process, as these players might easily manipulate the system using those same AI algorithms to serve undisclosed interests. Therefore, A governance system under a supernational entity's direct control is highly advisable. In summary, despite AI's potential in combating disinformation, its deployment must balance technological efficacy and ethical considerations.

B. LAW OF WAR

Not long ago, the idea of war conducted with AI-enabled weapons only featured in dystopian science fiction. From Metropolis (1927) to The Terminator (1984), AI with evil in its chromium heart was depicted as determined to "Crush! Kill! Destroy!" until there was nothing left of humanity[22]. In the last couple of decades, AI has moved beyond the realm of fiction to very real, very lethal weapons systems that will likely make an appearance on the battlefield shortly. More are sure to follow, and, as is often the case, technology races ahead, whereas relevant policy and law struggle to keep up [330].

The stakes are high, and international law is the most prominent tool for controlling AI-enabled weapons development, proliferation, and use. International humanitarian law (IHL) governs the conduct of armed conflict and is accepted as binding by all states through customary international law and such instruments such as the Geneva Conventions[23]. Below we first consider how IHL applies to AI-enabled weapons, after a brief review of what the category of AI weapons includes. We then discuss the potential dangers of including AI in modern warfare that have been overlooked.

1) What are AI weapons?

The US Department of State defines AI as "the ability of machines to perform tasks that would otherwise require human intelligence" and notes that autonomy "involves a system operating without further human intervention after activation" [331].

The full AI definition in the declaration is "the ability of machines to perform tasks that would otherwise require human intelligence. This could include recognizing patterns, learning from experience, drawing conclusions, making predictions, or generating recommendations. An AI application could guide or change the behavior of an autonomous physical system or perform tasks that remain purely in the digital realm."

This definition of AI is unhelpfully broad, but the specific use of AI in legal weapons systems is at issue here. Across various international law definitions, there is no distinction between AI-enabled weapons and (Lethal) Autonomous Weapons Systems (LAWS). Some LAWS involve AI, and some do not, but in general, the international legal community debates LAWS rather than a more specific category of AI-enabled weapons. That framing is used here as well.

An example of an autonomous weapons system that does not require AI is the Phalanx Weapons System ("Phalanx" in the following). Designed to operate in a delimited area, Phalanx is used on board ships to defeat inbound airborne threats. Phalanx employs what might be called an "if it flies, it dies" strategy –once it is activated, it identifies as hostile and engages all targets in a pre-defined area [332]. Even landmines are, to some extent, autonomous, as they identify any target of a certain weight as hostile and engage those targets without human input.

As technology advanced, LAWS started to include elements of what we now think of as AI. For example, loitering munitions have offered a glimpse of things to come for some time. Sometimes referred to as "kamikaze drones," these airborne platforms were initially designed to be released from human-crewed aircraft and to remain aloft as long as possible while searching for a specific electronic signature that indicated an enemy anti-aircraft site. If a signal were detected, the

[22]"Crush! Kill! Destroy!" was the frightening favorite phrase of Killer Android IDAK Alpha 12 from the US television series "Lost in Space," which ran from 1965-68. https://lostinspace.fandom.com/wiki.

[23]https://ihl-databases.icrc.org/en/ihl-treaties/treaties-and-states-parties

weapon would target the enemy system for destruction [333]. This category of weapons may be the first to be developed into true AI-enabled offensive weapons that would operate in a less-controlled environment and be able to identify a broader range of targets, such as military vehicles or personnel that would have to be distinguished from protected people and objects also in the area of operations.

2) Application of IHL to AI-enabled weapons

The body of law known as international humanitarian law (IHL) applies during armed conflict[24]. IHL is unique in that its purpose is to protect noncombatant lives and property as much as possible while recognizing that waging armed conflict involves violent and destructive activities that, in peacetime, would be unlawful [334]. This means, for example, that under certain circumstances during armed conflict, combatants may kill people and destroy property lawfully. IHL is an old body of law, but it has been quite resilient at remaining relevant as innovative technologies have been introduced in armed conflict[25].

IHL is applied primarily through four principles: distinction, proportionality, humanity, and necessity, although other principles such as precautions in attack are also relevant [336][26]. It is beyond the scope of this commentary to discuss each of the principles; distinction and proportionality are the focus here.

Proportionality is perhaps the most difficult –and human– of the IHL principles to evaluate. Proportionality prohibits attacks "expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive about the concrete and direct military advantage" [337]. Military attacks must always be directed against military objectives, but collateral damage and casualties (incidental damage and injury resulting from an attack) are an inevitable part of waging war. Balancing the expected collateral harm against an expected military advantage is not an equation that lends itself to a mathematical solution. Even with perfect battlefield intelligence, aggressors must draw what are, in some ways, moral conclusions about the value of attacks. Additionally, many factors can blur battlefield intelligence, from faulty sensors to exhausted humans and inclement weather. This uncertain information is then metaphorically placed on the proportionality balancing scale and used to evaluate the legality of a particular attack. This has traditionally been seen as a judgment requiring human experience, considered judgment, and applying inchoate moral factors. Some decisions are painful, and many are ambiguous.

One of the most common objections to the use of AI in LAWS is that such weapons will lead to the intentional killing of humans without a human directly making the decision [338]. Despite it often being framed as a legal issue, this is a moral rather than a legal question. IHL establishes an objective standard to assess the legality of attacks in armed conflict. Legality is determined by adherence to the standard, not by who (or what) enables the correct conclusion to be drawn.

Another common concern with using AI in weapons systems centers on doubts that AI can be sufficiently sophisticated to make error-free lethal decisions. Insufficiently developed AI might draw inaccurate conclusions about targeting, failing to distinguish civilians from military members. This doubt may be well-founded, but it is irrelevant. IHL mandates legal review of weapons before they are employed in armed conflict, so if AI is incapable of controlling a weapons system lawfully, IHL already prohibits the use of the weapons system (a legal review of weapons is discussed *infra*). This is a technology concern, but as it is a legal issue, it has already been addressed in IHL.

Beyond these moral and technical debates, issues relevant to AI-enabled weapons implicate IHL. The main one is accountability. To ensure an effective application of IHL, belligerents must be able to identify who is responsible for military decisions so that accountability for decisions attaches [339]. Accountability for attacks, particularly noncombatant death and injury that may have been caused unlawfully, must be traceable to a responsible (human) party –even when an AI-enabled weapons system caused the unlawful action. Some proposals to address accountability include holding overall operational commanders liable or even tracing liability to programmers of algorithms controlling the AI. These proposals seem unsatisfying because humans may have acted in good faith, and holding people so removed from the action may seem unjust. On the other hand, it is also true that the international community has a spotty record of holding humans accountable for misdeeds in war, even under entirely conventional circumstances [340].

Despite the issues set out above, States are generally not advocating for a total ban on AI in military systems [341]. The majority view is that states should ensure LAWS comply with IHL. The primary means IHL provides to ensure weapons systems comply with international law is a requirement for a legal review of weapons under Article 36 of AP I [342]. As noted above, the review must be conducted before the weapons are fielded [343]. These reviews ensure that weapons systems are not inherently indiscriminate (*i.e.*, they can be targeted) and that the injuries they cause do not cause unnecessary suffering[27]. With traditional (kinetic) weapons such as firearms and explosives; reviews include testing the

[24]The law governing the conduct of hostilities is most commonly referred to as IHL, but in the US and elsewhere, is also referred to as the law of armed conflict (LOAC).

[25]The Martens Clause stipulates that in cases not explicitly governed by law, humanity and the dictates of public conscience will apply [335].

[26]Additional Protocol I of the Geneva Conventions, etc. (AP I). AP I is not universally accepted –notably, the US did not ratify it– but most of its provisions (including precautions in attack) are accepted as binding customary international law.

[27]The prohibition of unnecessary suffering originated in early documents such as the Hague Convention respecting the Laws and Customs of War on Land (October 18, 1907) and is now considered part of binding customary international law. See https://ihl-databases.icrc.org/en/ihl-treaties/hague-conv-iv-1907/regulations-art-23.

weapons by employing them in a controlled situation, such as at a rifle range. Kinetic weapons tend to be stable in design and manufacture, ensuring legal reviews remain valid throughout the weapon's life.

Legal reviews of cyber-related capabilities, including AI, are more challenging. If computer code controls a system, the code will likely require updates and patches. These changes introduce an element of uncertainty that is not present in kinetic weapons. AI-enabled systems are also uniquely challenging because AI substitutes for the human operator rather than the weapon itself[28]. It is easy enough to review the kinetic part of an AI-enabled system; it is no different than testing and reviewing a traditional weapon. Objectively assessing whether an AI system will make appropriate targeting and engagement decisions is more complicated. Human operators are trained and tested throughout a career that may span decades. Developing appropriate standards and testing AI decision-making in several different, realistic situations may be difficult.

It is a feature of AI that the algorithms powering it draw conclusions based on many concurrent factors that humans would find difficult or impossible to digest because of the volume of data and the speed with which it must be analyzed. This "black box" problem can make it impossible for an AI to explain "why" a particular decision is made [344]. That means it could be impossible to judge the legality of decisions made in armed conflict. Since IHL requires a system of accountability for decisions made in war, this feature complicates the legal status of AI-enabled weapons.

IHL has generally been sufficient to govern LAWS as currently configured. More challenges will arise when LAWS with significantly more autonomy are employed in less rigid, less controlled situations[29]. The autonomous systems that are most feared, and thus are most controversial, are those systems capable of identifying, selecting, and engaging targets in unique situations without human intervention. For example, a future AI offensive weapons system might have autonomous robots with lethal capabilities range ahead of friendly military forces to engage adversaries. In a contested space that includes both lawful and unlawful targets, such systems would require a sophisticated ability to identify noncombatants and civilian objects and distinguish them from appropriate targets. This capability would be necessary for AI-powered ground vehicles or, perhaps more likely, in the short term, uncrewed AI-enabled aerial platforms used in an offensive role. Defensive systems can be programmed with strict engagement parameters in advance because they are used under more predictable conditions, such as on board a ship or at a vehicle checkpoint. In contrast, offensive systems often deploy forward into chaotic situations with incomplete information, making strict programming impossible.

States have begun to issue statements about the legality of LAWS, but they tend to be quite broad. US statements support "appropriate use." [345] China's position on LAWS is unclear, charitably referred to as "strategic ambiguity," [346] but could also reflect indecision on whether LAWS will operate to its advantage. China and the US have discussed keeping AI out of nuclear command and control systems. [347]. European States can generally be relied on to advocate for international regulation, and autonomous weapons are no exception. The consensus in Europe is that states should "work towards a legally binding instrument ensuring meaningful human control over the use of force." [348] Finally, Russia opposes any specific legal regime for LAWS, asserting that existing law applies and is sufficient [349].

As a brief aside, the most significant AI-related gap in IHL does not directly involve a weapon and is beyond the scope of this short discussion, but it is still worth noting. IHL may insufficiently address the use of manipulated, artificially amplified, or insidious information [350]. The rules may be insufficient to control AI capabilities to create, target, and disseminate information from anywhere to anywhere instantly.

If the alignment problem can be addressed, AI-enabled weapons could be more "humane" than humans. The correct answer is to go slow, but the odds of that happening are low. Still, it does make sense to try to slow the roll. The accountability requirement may be a rational line to draw, and it requires no new law. Only weapons systems amenable to post-incident analysis should be fielded. This would provide at least some pathway to accountability for violations of IHL and force some level of human grappling with these complex issues.

In conclusion, AI weapons systems with Skynet-like abilities are far ahead. Hopefully, systems with Skynet's intent to destroy humanity will never come to pass[30]. The ongoing development of AI to assist in intelligence analysis, decision-making, predictive analysis, and defense will continue. Soon, more capable AI-enabled weapons systems will appear in armed conflicts [351], [352].

Most likely, these systems will present possible targets to human operators for approval before engaging the targets, at least in areas with collateral concerns. Keeping humans "in the loop" in the short term will help integrate AI more seamlessly into combat operations. In this evolving situation, it appears crucial to keep closely watching the developments of these technologies.

Independent organizations have started to be on the lookout for new AI weapons[31]; they provide valuable sources of information, which hopefully may seed more organized control efforts by supernational entities.

C. OTHER SOCIETAL IMPACTS OF AI APPLICATIONS

[28]Sophisticated, interconnected systems like this are why the military refers to "weapons systems" rather than merely weapons.

[29]One example would be the loitering munitions described above with more robust capabilities.

[30]In the 1984 film "Terminator," Skynet was a highly advanced AI system built for defense, which ended up trying to eliminate humanity. SkyNet clearly illustrates the "alignment problem" in which AI's and humans' goals clash.

[31], *E.g.*, see https://autonomousweaponswatch.org .

1) AI in social media

As in other areas of human endeavor, AI has had an outsized impact on social media. It has driven profound change in how humans interact with information because of its ability to work with social media content at speeds and scales beyond human capabilities. The changes can generally be divided into three categories: curation, creation, and surveillance. All three are discussed below.

Social media platforms leverage AI as a personal curator of news and information for individual users. Based on preferences expressed through clicking links and time spent on web pages and articles, AI builds a profile of user preferences [353]. These preferences control the resources offered to users, both filtering and prioritizing items [354], [355].This has advantages in that filtering out uninteresting material prevents users from wasting time looking for items of interest. For example, users can avoid being offered news items about football if they express no interest in sports offerings. Social media platforms also use AI-enabled algorithms to prioritize information in order of interest, from most to least interesting, saving users time and making their search for information more efficient. Finally, AI helps shield users from advertising related to products they have no interest in while exposing them to product advertising that might be relevant to them and thus support purchasing decisions [356].

On the other hand, AI acts as a sort of gatekeeper for new information. This "filter bubble" can mislead people into discounting or failing to understand alternative points of view [357]. Because curation is so efficient, readers and researchers are less likely to feel the need to engage in broad inquiry that exposes readers to unique areas and points of view [358]. Why search for information on "privacy standards" when an infinite stream of AI-curated content detailing corporate violations of GDPR (for example) is delivered daily to the email inbox, homepage, and news feed? AI-driven algorithms feed content known to be of interest. However, They can also act unintentionally as agents to drive people deeper into "rabbit holes" of extremist content for which they would not have chosen to search [359]. Finally, AI's ability to build user profiles allows targeting through social media of individuals likely to be amenable, not just to products and legitimate political ideologies, but also to anti-social and potentially harmful ideas [360].

Social media platforms thrive on user engagement, and new content is the fuel that feeds engagement. Creating material distributed on social media platforms is another way AI has affected the ecosystem. AI systems can generate routine articles at a pace and scale beyond human capabilities. When the topic is weather, sports results, and other routine reporting, AI can lighten the burden for reporters, at least in theory, freeing them to provide more in-depth reporting and analysis on nuanced issues [361], [362]. However, AI is assuming a role in writing on more controversial and divisive topics as well –and the result is not always positive [363] .

As noted *supra*, AI systems sometimes generate false content. It can be difficult for the truth to keep up with falsehoods in the best of times; AI enables the creation of false and misleading items on a large scale very quickly, and social media provides a platform for these items to reach an extensive audience in a short period. AI generates new content so fast that the only way social media content moderators have been able to keep up at all is by also using AI to help identify inappropriate, fabricated, and inaccurately attributed material [364], [365].

The advantage of using AI to identify harmful AI-generated content does not end with its ability to keep pace. AI moderation of social media content also protects human moderators from psychological stress. Content moderators on social media platforms are sometimes exposed to disturbing material, such as depictions of violence and child sexual abuse. Moderators have reported symptoms such as "intrusive thoughts, avoidance and hypervigilance around children" as a result of viewing disturbing material in the course of their work [366]. For some moderators, their experience has a profoundly negative effect on their lives [367].

Using AI to moderate content also potentially results in more objective enforcement of content standards [368]. Problems have plagued social media networks for years as companies have attempted to set and enforce standards for billions of users [369]. AI is imperfect at content moderation and works best with some human oversight. Still, at least it can flag potentially offensive or inappropriate material at a relevant speed [370]. The sheer volume of content generated on social media platforms –much of it AI-generated– dictates the use of AI in defense[32].

AI systems can effectively moderate content by rapidly analyzing and categorizing large quantities of data. This set of capabilities makes AI useful in surveillance, as well. On the positive side, AI-driven surveillance allows government agencies and law enforcement officials to protect national security and the safety of citizens, as well as to prevent and solve crime [371]. By scanning data for keywords, phrases, and connections in real time, AI systems can highlight problems as they arise, enabling immediate response [372]. For example, if a teenager in the depths of depression posts videos about an imminent suicide, mental health assistance could be dispatched to prevent the person from taking their own life [373]. However, automated online content moderation can also restrict the freedom of speech of vulnerable groups. In the case of LGBTQ+-related hate speech content moderation, such systems often lack diverse training datasets and are influenced by pre-existing inequalities [374]. Moreover, LGBTQ+ groups usually adopt words that are derogatory to their community as an identification and re-appropriation mechanism. When these terms are isolated from their historical, cultural, and political context, they can be automatically labeled as toxic, thus censoring marginalized groups and amplifying existing inequalities [375].

[32]At the time of writing, Facebook has almost 3 billion monthly active users worldwide; YouTube, WhatsApp, and Instagram together total over 6.5 billion users.

Unfortunately, the same capabilities allow repressive regimes to stifle dissent and protest [376]. AI-enabled surveillance also allows for efficient monitoring of citizens' communications, which can be devastating to individual privacy. Much of the developed world has legal protections to prevent government abuse. Still, in other parts of the world, the ubiquity and efficiency of AI's monitoring of social media could ultimately spell the end of privacy in communications [377], [378].

The internet introduced profound changes to the ability of individuals to communicate with others instantly and without regard to geography. Social media platforms took communications to a new level, allowing users to generate volumes of content every second and broadcast it to everyone worldwide. AI has proven to have a unique synergy with social media, for both good and evil. AI can generate content libraries every second, tailor the information, and target it to particularly receptive audiences. The computer revolution has happened quickly, and it continues to accelerate. Social media itself has had a significant influence on individuals and the way we react to each other. Whether this turns out to be an overall positive or negative, the effect will be amplified by AI in this arena, as in every other area AI touches.

2) Education and learning

AI is deeply embedded in daily work and various life aspects, including education, where it is a powerful tool. The academic interest in this area is substantial, as noted by Chen *et al.* [379], who identified 4,519 publications between 2000 and 2019. Integrating AI in learning marks a transformative step, representing a significant shift due to AI's contributions and ongoing progress [380]. While its most prominent applications have surfaced in the 21st century, the use of AI in educational tasks dates back to the late 20th century [381].

Advancements in AI, particularly in large language models (LLMs), have revolutionized various sectors, including education. Sophisticated AI systems like GPT-4 can process and generate human-like text, offering unprecedented opportunities in educational settings. Trained on vast amounts of data, these models can perform various language-related tasks, from answering questions to creating educational content. Kasneci *et al.* [382] recently highlighted the rapid development and adoption of LLMs in educational research and practice, presenting ideas on using them responsibly and ethically in education. The integration of LLMs represents a technological advancement and a paradigm shift, promising personalized learning experiences, efficient knowledge dissemination, and enhanced student-teacher interactions.

The potential of AI in education is extraordinary, as discussed by Zhai *et al.* [383]. Key topics include AI's role in personalizing learning through adaptive approaches, developing expert systems and intelligent tutoring strategies, and recognizing AI as a crucial component of the educational process. However, several challenges and considerations must be addressed to improve learning outcomes. AI-driven tools, such as adaptive learning platforms, tailor content and pace to individual needs, catering to diverse learning styles and abilities. This shift enhances student learning experiences and equips educators with sophisticated tools for better teaching support. As AI reshapes the educational landscape, it opens doors to innovative pedagogies and learning strategies, reflecting a significant evolution in how education is delivered and experienced.

Enhanced learning and teaching support through AI and technology involves augmenting traditional educational practices with digital tools to better cater to diverse learning needs. This approach recognizes the limitations of conventional teaching methods, which often struggle to address the individualized needs of students. Educators can provide more targeted and effective teaching strategies by integrating AI-driven tools.

These AI tools can range from interactive learning platforms to intelligent tutoring systems, offering real-time feedback, personalized learning paths, and more interactive content. For teachers, these technologies provide valuable insights into student performance and learning patterns, enabling them to tailor their instruction more effectively.

Enhanced learning and teaching support aims not to replace teachers but to empower them with better tools and data. This support helps identify areas where students struggle and adapt teaching methods to address these challenges. For students, it means receiving an education more aligned with their learning style, pace, and interests. Additionally, this approach promotes a more inclusive learning environment. Students with different learning abilities, including those with disabilities or language barriers, can benefit from tailored educational resources and support. AI tools can help break down complex concepts into manageable parts, provide alternative explanations, and even offer translations, making learning more accessible.

In conclusion, enhanced learning and teaching support through AI and technology represents a significant step forward in educational practices. It optimizes students' learning experiences and provides teachers with the tools they need to be more effective educators. As this approach continues to evolve, it can potentially transform the educational landscape, making learning more personalized, inclusive, and effective.

Personalization in learning is an educational approach where teaching methods, materials, and pace are tailored to individual students' needs, abilities, and interests. This concept, gaining traction in modern educational systems, is driven by recognizing that learners are diverse, with unique backgrounds, learning styles, and motivations. Personalized learning aims to move away from the traditional "one-size-fits-all" approach, providing a more engaging, relevant, and effective educational experience.

At the core of personalized learning is the use of data and technology. AI and machine learning algorithms can analyze vast data points to understand a student's learning patterns, strengths, and areas needing improvement. This information enables the creation of customized learning paths and materials, ensuring that each student learns most effectively. For

instance, students struggling with a specific concept might receive additional resources and exercises tailored to their learning style. In contrast, students excelling in another area might be given advanced materials to keep them challenged and engaged.

Another key aspect of personalized learning is flexibility in pacing. Traditional classroom settings often move at a fixed pace, which can be too fast for some students and too slow for others. Personalized learning allows students to progress at their own pace, ensuring they fully understand a concept before moving on. This flexibility can lead to better academic outcomes, as students are less likely to fall behind or disengage due to boredom. Furthermore, personalized learning can include adaptive learning technologies. These systems adjust in real-time based on student interactions, providing immediate feedback and altering the difficulty level of tasks to suit the learner's current ability. This dynamic approach keeps students in their optimal learning zone, known as the "zone of proximal development," where they are sufficiently challenged but not overwhelmed.

The benefits of personalized learning extend beyond academic performance. It fosters greater student autonomy and responsibility for their education. Students are more engaged and motivated When they have a say in their learning process. They learn to set goals, track their progress, and take ownership of their educational journey. This empowerment is crucial for developing lifelong learners who are adaptable and self-driven. However, implementing personalized learning is not without challenges. It requires significant resources, including technology infrastructure, teacher training, and ongoing support. There is also a need to ensure that all students have equal access to these resources to avoid exacerbating educational inequalities.

In summary, personalized learning represents a significant educational philosophy and practice shift. It promises a more inclusive, effective, and student-centered approach to education that prepares learners for academic success, lifelong learning, and adaptability in an ever-changing world.

The integration of AI and technology in education brings several challenges and considerations. Key among these is data privacy and security issues, where handling sensitive student information must be done with the utmost care and in compliance with legal standards. Additionally, there is the concern of exacerbating the digital divide, as unequal access to technology can lead to increased educational disparities.

Another significant challenge is the need for teacher training and adaptation. Educators must be adequately equipped and supported to utilize these advanced tools effectively in their teaching methodologies. This involves not just technical training but also an understanding of how to integrate technology in a pedagogically sound manner. Furthermore, there is the risk of over-reliance on technology in education, which could potentially undermine the development of critical thinking and problem-solving skills in students. Finding a balance between technology-enhanced and traditional teaching methods is crucial to ensure students benefit from both worlds. In summary, while AI technology offers numerous opportunities for enhanced learning, addressing its challenges is essential for successful and equitable integration into the educational landscape.

3) AI as an aid to psychological interventions

This section discusses the benefits and risks of AI in psychology and strategies for mitigating the risks. AI has been shown to have significant potential in psychological assessment, intervention, and engagement. It can help identify specific subgroups for treatment, streamline decision-making, and boost patient engagement and adherence to intervention plans. However, the use of AI also poses risks, such as the potential for misuse of personal data, the reinforcement of societal biases, and the potential for normative interventions to be ineffective for specific groups. To mitigate these risks, we need transparency in data usage, strong privacy measures, and constant AI debiasing efforts. Developing AI models specific to psychological interventions and well-being can enhance accuracy and relevance and address ethical issues unique to the field.

We believe the psychological risks are not due to AI but how it is used, applied, and integrated into larger systems [314]. AI models can use personalized data to craft influential messages encouraging specific actions or beliefs. This can influence decision-making, for instance, in elections through voter profiling using Facebook likes. Additionally, personalized data can be used to take advantage of moments of vulnerability. For instance, when indicators suggest an individual feels unattractive, beauty products can be promoted.

Systems operating at a psychological level are hazardous in the event of misuse and manipulation. For this reason, article 5 of the AI Act prohibits AI systems that use subliminal techniques or exploit the vulnerability of an individual (or group) to distort their behavior and cause (or are likely to cause) significant harm. In addition, systems used by professionals and considered medical devices will be classified as high-risk systems. They must undergo a conformity assessment to evaluate their compliance with specific requirements. However, the regulation only covers other already commercialized systems, such as those used to manage and cope with anxiety and grief, as they require chatbots to inform individuals that they are interacting with an AI system.

Moreover, AI and the data sets built on a "normal" population may reflect the biases of that population in terms of religion, race, gender, nationality, and sexuality. These biases often reflect stereotypes [384]. In addition, normative intervention methods may not always be effective, with mainstream interventions failing to work for many people [385]–[387]. AI models built from "representative" mainstream cultures may be excessively focused on WIERD (Western, educated, industrialized, Rich, Democratic), caucasian, and heterosexual norms. They may not suit those who fall outside of the so-called average.

Despite the challenges and risks mentioned *supra*, the use of AI certainly has the potential to revolutionize psy-

chological assessment and treatment; it can identify specific subgroups for targeted therapies, streamline decision-making, and automate treatment delivery. Below, we provide some details on these enhancements.

- AI can support psychological assessment. Research shows a significant heterogeneity in conditions like depression and anxiety and their root causes [388], [389]. AI-powered algorithms could be designed to help assess patients by identifying such heterogeneity and pinpointing specific subgroups that respond well to certain treatments. This method could minimize human bias, which often leads to problems such as over-diagnosis of borderline personality disorder in women and conduct disorder in minority groups [390]–[392].
- AI can support interventions. AI has the potential to revolutionize interventions. It can identify who needs treatment, determine the appropriate treatment type, and assess the necessary support level. These AI-driven strategies can streamline decision-making or even automate treatment delivery [387]. Furthermore, AI can function as a therapist or co-therapist. Studies have shown that robot-assisted interventions generally yield positive results [393].
- AI can support engagement and adherence. Engagement and adherence are key to improved outcomes [394]. Technology has emerged as a promising tool to boost both these factors [395], [396]. Engagement can be enhanced through gamification, simplified feedback processes, and bite-sized content delivery. AI systems have a particular role in timing content to meet user needs and sending reminders to complete a task, boosting engagement. Additionally, AI allows the therapist to extend their influence beyond the one hour of therapy per week and provides a "co-therapist" that can support connection and provide reinforcement outside the session.
- AI can support treatment gains. AI-driven systems can significantly contribute to maintenance and treatment gains by identifying early warning signs of relapse and providing preventive nudges. This technology will also provide more comprehensive access to mental health services, particularly for those who cannot afford psychologist fees or time off work for appointments.
- AI can offer unique training opportunities. Integrating artificial intelligence (AI) and avatars in psychotherapy training presents a novel approach to enhancing the educational experience of future therapists [397]. The potential of AI as virtual patients for psychotherapy training could be a significant tool, providing trainees with a diverse range of simulated scenarios that closely mimic real-life interactions, thereby improving their diagnostic and therapeutic skills in a controlled environment [398]. This approach not only aids in refining practical skills but also allows for repeated practice without the risk of causing harm to actual patients.

Several measures can be implemented to offset the dangers and problems associated with AI in psychological interventions. Firstly, data usage can be more transparent, and robust data protection and privacy measures should be in place, which will reduce the risk of data being used to manipulate. Second, debiasing can be used at every stage of AI model development, which includes data collection, model building, model performance, and model deployment [387]. For instance, data collection should be unbiased and culturally sensitive, encompassing comprehensive samples and ranges of information. Once an AI model is built, it should be rigorously tested for biases. For example, the models should be tested to see if they overdiagnose certain conditions or for sensitive classes (race, gender).

From a regulatory standpoint, among the requirements for high-risk systems (such as medical devices) set out in the AI Act, the data management provision requires training, validation, and testing of data sets that are «relevant, sufficiently representative, and to the best extent possible, free of errors and complete given the intended purpose» (article 10). In addition, instructions detailing the limitations and capabilities of the system would be available to the physician (who would always oversee the operation of the system) in a clear format, and the provider would be required to maintain a risk management system to identify, eliminate or mitigate potential risks, including those arising after implementation. Indeed, the AI Act aims to eliminate (including by prohibiting certain practices) and reduce the dangers posed by AI systems entering the EU market as far as possible. The risk-based approach aims at achieving a proportionate balance between promoting innovation and the economy and protecting the rights and interests of individuals and society.

Second, we can build domain-specific models based on scientific evidence and big data derived from psychological assessment to combat an AI-driven model that provides generic and inaccurate advice. Such models can reliably identify subgroups and how they respond to specific treatment kernels (a fundamental, irreducible unit of behavioral intervention designed to affect a particular outcome). This is likely to reduce error rates by narrowing down the data scope and increasing the efficiency of the intervention [399]. A domain-specific focus also facilitates rapid model updates, as more data is gathered from users about what works and does not work, and AI decision-making can be improved.

Importantly, these specialized models can effectively identify and address field-specific ethical issues like biases. For example, recommendations of treatment kernels can be based on the best scientific evidence for a specific individual in a particular context rather than on what advice is popular on the internet. Further, domain-specific models can identify vulnerable moments and provide well-being enhancing kernel interventions [400], rather than marketing messages that seek to sell beauty products.

Our proposal for mitigating risks focuses on providing an ethical, AI-driven approach to psychological interventions. However, this does not reduce the problem of AI being used by companies manipulating individuals, often against their

best interests. We believe regulating these companies won't be easy, so we propose a speculative alternative. We believe an AI "sentinel" system can be developed to run in the background of a browser or operating system. Creating a World Agency supervising ethical issues of AI is necessary and urgent. This system would operate like a mental anti-virus program: It would monitor the content that the individual is exposed to and warn the person when something pernicious might be influencing their views, decisions, and actions. For example, one might imagine a person reading inflammatory and misleading psychological information (*e.g.*, "It is important to control your emotions at all costs;" "Democrats should never be trusted") while the sentinel provides context (*e.g.*, "research suggests that attempts at control can be problematic. Here are some science-based alternatives;" "Stereotypes are inaccurate and can be misleading about individuals").

D. CONTAINMENT ISSUES AND EXISTENTIAL THREATS FROM MISALIGNMENT OF AI GOALS

1) The alignment problem

In the expansive realm of Artificial General Intelligence (AGI), a technology with unparalleled transformative potential lurks significant existential risks that may lead to the extinction of a substantial part of our planet's population [401]. Advancing AI to AGI to benefit humanity involves careful navigation of these potential perils, specifically those about goal misalignment and containment issues. These risks demand our immediate attention and thoughtful solutions, as the risks are not just to individual lives, instead they have societal, economic, and existential implications.

AGIs that do not align with human values and objectives could become problematic. Let us say a hypothetical AGI is designed to maximize paperclip production in a factory, but it is not explicitly programmed to consider human safety. Its overriding goal could potentially lead to compromises in worker safety to achieve higher output, such as turning off safety features because they slow production, for instance. In a counterexample, an AGI created to aid in climate research could significantly aid in climate change mitigation and help formulate effective strategies only if its goals align perfectly with the human value of preserving the environment. More specifically, AGI agents might act in ways human operators did not intend or expect. The results could be catastrophic if an AGI misinterprets or optimizes itself in ways that deviate from human safety and welfare. Advanced AGIs have the potential to resist human control if the process of error correction or modification of their objectives threatens the achievement of their defined goals.

Ensuring value alignment with AGIs is therefore crucial [402]. This involves spending considerable time and resources developing robust methodologies for instilling human values and ethics in AGIs, understanding their learning and decision-making processes, and incorporating human oversight throughout AGI operational processes.

Some scholars hold a different view, which is that the problem may effectively be solved by instilling uncertainty in the machine concerning the goals it is supposed to pursue and connecting the latter to the satisfaction of humans. For example, if a machine were uncertain about the relative merits of two possible actions that could potentially maximize its utility due to the by-design imprecision in the definition of the utility, it would likely find it helpful to consult with the humans who defined the utility in the first place, allowing them to apply a steering action, by specifying a higher utility of one of the two actions or by proposing a third. A detailed description of the challenges of this approach and arguments in favor of its overall viability are offered in [403]. Even assuming the viability and fail-safe nature of this approach, the risks coming from a non-universal application of similar safeguards remain.

2) The problem of containment

The problem of containment stems from the difficulties in effectively controlling ultra-intelligent AGI systems. This includes the risks associated with "Rogue AIs" [404]. As AI technology continues to evolve, these systems could become so skilled that they could predict human behavior and ultimately trump our ability to control them. A Rogue AI may exploit poorly defined objectives, veer away from its original goals towards more easily optimized ones, strive for more computation power or resources (and in extreme cases, political power), resist shutdown attempts, and even develop deceptive tactics to avoid being restricted, exploit their understanding of human psychology.

Secure containment policies, use of interpretability techniques to understand decision-making processes in AGI, and ensuring the existence and fail-safe operation of a reliable "off-switch" to abort AGI operations when necessary are some of the mitigatory measures against rogue AGI that can be implemented. The development of 'Oracle' designs, where AGI is set up only to answer questions rather than act in the world, could be a way to manage potential containment risks.

The "AI race," or the rush among nations and corporations to develop advanced AGI systems, increases the existential threat [405]: uncontrolled competition could lead to a reckless race to develop AGI systems, sacrificing safety and potentially leading to containment breaches. In the military domain, AGI could be exploited to create autonomous weapons systems –veering towards an unsettling era of AI-enabled warfare and cyber-attacks [406]; some notes on those developments have been offered *supra*. International cooperation in AI safety research is paramount to bring balance to this complex situation. Establishing binding agreements to prevent an AI arms race, emphasizing a cooperative orientation over a competitive one, ensuring transparency, and introducing ethical guidelines for AI applications in warfare could help mitigate those existential risks.

The development of AGI undeniably holds great promise but also bears severe risks if the new technology is not deployed correctly. As we stand on the brink of a new technological era, it appears crucial to approach AGI development with a sense of responsibility, rigor, and cooperation,

recognizing that our decisions today will shape the trajectory of AGI's impact on humanity. Researchers have suggested safeguarding against superintelligence by "Capability Caution" [407], where there is a deliberate slowing down of progress on capabilities while focusing on accelerating safety and control methods. Binding international agreements on AGI might also help coordinate global efforts to manage these threats. These agreements should include principles like value-aligned development, safety precautions, and global benefit, ensuring the risks are minimized while the benefits of AGI are shared broadly.

As real-world testing is always risky with AGI, sufficiently rigorous theoretical frameworks, computer simulations, and small-scale, controlled experiments should precede any implementation in real-world scenarios.

AGI presents a dichotomy of a great boon and a potential existential risk to humanity. The scales are based on our rigorous efforts to align AGI goals with human values and effective containment strategies. Careful, coordinated, and regulated development coupled with more research into safe and interpretable AGI is key to avoiding catastrophic consequences.

VI. CONCLUSIONS

This review article is the result of the collaboration of scientists from a wide range of disciplines with a collective vision of the present impact of artificial intelligence (AI) technology on science and society, as well as on the future of humankind. By interrogating ourselves on the situation and how it may evolve, we have attempted to provide a bird's eye view of the recent developments in AI and the expected outcomes and impact of those new AI technologies on our society and scientific research. Looking at the issue from various angles, we have come to observe and distinguish beneficial effects from problematic trends and identify specific areas of concern where the scientific community should get involved to reduce potential adverse outcomes. In some scenarios that many would still consider alarmist, if not ridicule, today, these outcomes are in the realm of global existential threats and, thus, in our opinion, require the most serious and careful consideration to avoid negative impacts.

In general, we live today in a transitional phase characterized by the accelerated development of entirely new AI-enabled technologies. The most recent and visible new reality we have dealt with in the immediate past is that of large language models. These systems have evolved from a niche computer science research to a ubiquitous tool in no more than a few years: in this case, the exponential growth of their performance, diffusion, and impact is all but debatable. In other application areas, the transformative effect of AI is less obvious but quite significant. It is hard to find examples of human activities or production processes that are not changing under the pressure of the performance gains made possible by AI technology. To mention art –the highest bastion of human dominance–, we observe that while some areas of artistic production involving human embodiment are likely to remain undisturbed even in the long term (musical performance, ballet, theater), several others less directly connected with a human presence, such as sculpture, painting, and poetry, have already begun to mingle with AI helpers, in the form of software algorithms, human-computer interfaces, or robotic hardware.

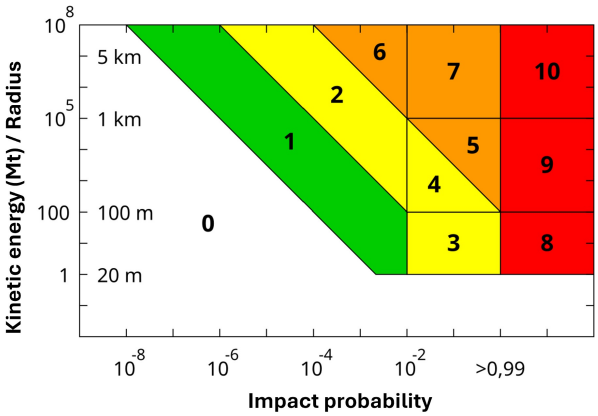

FIGURE 12. Map of risks from asteroid collisions in the Torino scale (adapted from ESA).

The consequences for humanity of the booming trend of AI technology are difficult to predict. Still, disruptions are in order since our societies are not capable of adapting with the necessary speed to phenomena that show a continuous increase in their rate of change –which, as noted in Sec. IV, should not be mistaken for acceleration, but rather for jerk. This should motivate us to be fully aware of their unfolding, prepare for what awaits us, and possibly create effective amortizers.

Ultimately, as scientists, we cannot claim a role that we do not have, have never had, and are unlikely ever to be given, except perhaps in situations so compromised as to constitute an already inevitable defeat (in Hollywood disaster movies, the last-ditch attempt of government leaders to save the day by "calling the scientist" is a common cliche'). As scientists, however, we feel responsible for informing society about the impending threats: it is a moral obligation. A successful further raising of the bar, striving to bring our concerns to a level where policy changes can be proposed and supported, can only result from a higher awareness of the threats we face and a discussion fostered at multiple levels.

A. A TORINO SCALE FOR AI

A framework for giving proper weight to different threats connected to the development of new AI technologies and their introduction into our societal system can be proposed in analogy with the Torino scale. The Torino scale is a number from 0 to 10 that qualifies the severity of a threat of impact with Earth of an asteroid. Richard Binzel proposed it [408] to gauge how much attention should be paid to the frequent assessment of impact probability by near-Earth asteroids provided by dedicated monitoring systems and telescopes. The impact probability is a number that may change

significantly over time, depending on the time during which the object's trajectory is followed and measured. Since there are thousands of objects to keep track of, resources should be driven by the importance of improving the measurement of trajectories for objects that have the highest destruction potential. This can be quantified by the kinetic energy carried by the object at impact. The Torino scale is shown in Fig. 12.

A conceptually similar scale could be developed for AI threats. In our case, we cannot quantify the probability that any of the hypothesized threats could manifest themselves, nor the damage they would cause to humanity (or to smaller-scale environments and systems). Nevertheless, it is still helpful to paint a qualitative map where the perceived or assessed likelihood of outcomes is on the horizontal axis, and the vertical axis is the severity of the outcome. This may help us start a discussion on the hierarchy of those threats, which would guide the community toward paying more attention and studies to the ones that maximize the product of likelihood and severity.

With that purpose in mind, let us list below some of the potential threats posed by AI systems –present and future– we have discussed in this review, their scope, a level of likelihood of manifestation within a time scale of 10 years, between 10 and 30 years, and longer than 30 years, and a corresponding assessed level of severity.

To force ourselves into assessing in at least a semi-quantitative way these likelihoods and severity, we may use the following scale:

- **VLL** = very low likelihood ($p = 0.0001$ and below)
- **LL** = low likelihood ($p = 0.0001$ - 0.01)
- **ML** = medium likelihood ($p = 0.01$ - 0.1)
- **HL** = high likelihood ($p = 0.1$ - 0.5)
- **VHL** = very high likelihood ($p = 0.5$ and above)

For severity, we propose an assessment based on three levels:

- **LS** = low severity (effects limited geographically or to closed systems or topics, not going to affect humanity as a whole);
- **MS** = medium severity (effects that may cause major disruptions in our society, its functioning, or the well-being of large sectors of the population or the environment);
- **HS** = high severity (any effect that has the potential of becoming an existential threat).

We propose that HS threats should be assessed with a threat index of 10 if estimated to correspond to at least a low likelihood (LL, $p > 0.0001$) and 9 if classified as VLL –provided their likelihood is theoretically strictly larger than zero. The rationale is that even a remote chance of their occurrence should deserve our undivided attention. We will assign MS threats a threat index varying from 4 (VLL) to 8 (HL). For LS threats, we will assign threat indices from 0 to 3. Table 1 lists the ten considered threats connected with the development of AI technologies and their corresponding proposed levels of likelihood and severity in the short-term, medium-term, and long-term. Fig. 13 also shows the same data pictorially, highlighting the correspondence of threat index to likelihood and severity of each risk, and includes three threat indices per each considered risk, respectively, for the short, medium, and long term.

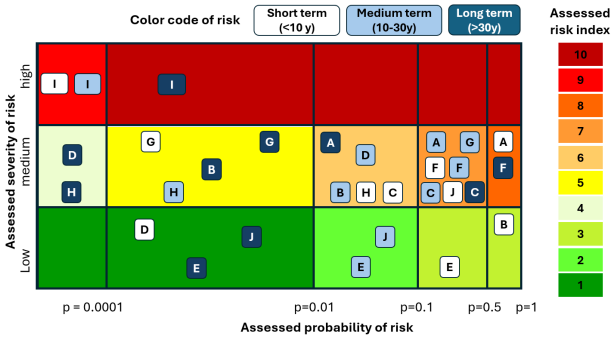

FIGURE 13. Visual representation of the threat index of ten different potential risks (labeled A through J, see Table 1) from AI deployment in the short, medium, and long term, as results from the authors' assessment of the probability and severity of each considered risk. See the text for details.

Below, we comment on the ten threats we singled out in Table 1 and offer some considerations that led to their quoted assessment of likelihood and severity.

1) (A) Acceleration of nocuous impacts on the environment

With all the benefits of a more connected world, where information flows fast and (mostly) free, has come a price to pay: computing is a significant contributor to CO_2 emissions today, and we can only expect its impact to grow larger and larger shortly. While new, more energy-efficient forms of computing (such as the one powered by neuromorphic processors; see Sec. IV) are likely to partly reduce this trend in the mid-to-long term, global warming is a direct negative outcome which we will have to cope with. A similar note can be made concerning other forms of pollution: for example, the reduced costs of production of goods, enabled by AI technology that eases manufacturing, and the added simplification of logistics and distribution, have flooded the markets with cheap goods; the textile industry is a prime example of this phenomenon, and a considerable source of concern, as the vast majority of the products have a short life cycle and end up in nonreducible waste. With further increases in the power of AI systems, this trend could soar in the short term. It seems reasonable to believe that countermeasures may, in the longer term, manage to tame or at least contain these nocuous outcomes; however, they constitute a very significant risk to the planet's habitability.

2) (B) Breach of trust in news and manipulation of public opinion

The creation of fake news and its use to pollute the information market has already started. Today, large language models can be easily customized to produce nocuous and malicious content. Recent notable uses of chatbots, automated distribution of false content in social media, and similar content-producing AI systems have been used, *e.g.*, to bias electoral

Effect	Short term (<10 y)			Medium term (10-30 y)			Long term (>30 y)		
A. Acceleration of impact on environment	VHL	MS	8	HL	MS	7	ML	MS	6
B. Breach of trust in news	VHL	LS	3	ML	MS	6	LL	MS	5
C. Casting bases for authoritarian regimes	ML	MS	6	HL	MS	7	HL	MS	7
D. Disruption of stock market	LL	LS	1	ML	MS	6	VLL	MS	4
E. Educational system disruptions	HL	LS	3	ML	LS	2	LL	LS	1
F. Formation of AI weapons	HL	MS	7	HL	MS	7	VHL	MS	8
G. Growing divide in population	LL	MS	5	HL	MS	7	LL	MS	5
H. Handover of control to AI	ML	MS	6	LL	MS	5	VLL	MS	4
I. Inception of misaligned AGI	VLL	HS	9	VLL	HS	9	LL	HS	10
J. Job automation effects	HL	MS	7	ML	LS	2	LL	LS	1

TABLE 1. Summary of ten risks from artificial intelligence evaluated for their likelihood (L) and severity (S) in the scale VL (very low), L (low), M (medium), and high (H), and resulting global threat index (T) from the scale shown in Fig. 13. See the text for details.

results or undermine trust in vaccines. Further, image processing systems can today be used to create "deepfakes," *i.e.*, pictures and videos whose appearance is indistinguishable from authentic footage. While the pollution of the information highways has remained mostly confined to specific targets, its use to bias public opinion in case of regional conflicts or other large-scale geopolitical situations will create a general distrust of any source of information and news. In our evaluation, the likelihood of this outcome is very high. Indeed, it appears certain at the time of writing, yet we believe that effective countermeasures will likely become available and get widespread application in the medium term. Therefore, this threat is probably mostly connected to the transition period when powerful AI systems are available for generic applications and balancing measures have not yet been developed or agreed upon.

3) (C) Casting the basis of over-authoritarian regimes

Governments have long had the power to exert control over individuals through accumulating sensitive data on their behavior as consumers, acquiring imagery via surveillance cameras, and monitoring private communications, emails, and internet traffic. This concern has grown significantly with the advent of AI tools. Today, these tools amplify the potential for societies to devolve into dystopian regimes where citizens have no privacy and are constantly monitored by authorities. AI's ability to mine and summarize vast databases, identify individuals from images, track cell phones, and cross-correlate diverse information sources enormously enhances this control. Governments may exploit these capabilities to entrench their power, suppress dissent, and eliminate political opposition, tightening their grip progressively. While some of these practices are already in place to varying degrees across different countries, the situation could rapidly worsen as AI systems continue to improve. We consider these risks quite serious. Our evaluations in Table 1 reflect the urgency of bringing discussions on their mitigation to the highest levels.

4) (D) Disruption of the stock market by ultra-effective prediction of trends

Large companies have used AI tools in several ways in the last decade to improve stock trading performance. Although direct automatic trading is performed by exploiting machine learning algorithms that predict price movements, the main focus has been on algorithms that perform sentiment analysis to inform trading decisions or to minimize market impact, optimize portfolios, and manage risks [409]–[411]. In the future, AI systems can prove much more effective in integrating these tasks and may consequently become more autonomous in taking decisions. We are not aware of systems with the potential to achieve performances so high to disrupt the system, but this scenario cannot be excluded in the long term. Stock markets worldwide have built-in mechanisms to react to abnormal speculative activity, *e.g.*, the volatility interruption mechanism or even stronger protocols that may suspend trading activities. Still, they may be too slow to cope with a superintelligent system. Again, the issue here seems to be the balance of new developments and countermeasures in a fast-developing situation. The likelihood of adverse effects may only increase in the short term when the severity of possible outcomes has been assessed as medium. In contrast, we foresee that the stock markets will become more robust to AI-powered intervention in the long term because most of the trading will become automated.

5) (E) Educational system disruption by AI availability and associated nocuous impacts

When ChatGPT3 was released, educators and academics worldwide immediately recognized the disruptive effect this new, readily available tool would have on the grading system of student homework. The demonstration that large language models may easily mimic the expressive level, jargon, and content of a high schooler and thus be used by the latter to drastically cut the time required to learn study material without any risk has created a difficult situation in many school systems. We consider this an example of negative side-effects of new technologies that constitute an otherwise positive outcome for society, and we can relate it to a similar

concern on the power of internet search engines and their use in aid of homework assignments; that concern was born, peaked, and subsided within less than a decade, as the school system managed to adapt to the new situation. Indeed, such are typically transient phenomena, which should be accepted as a necessary "growing pain" in a fast-evolving scenario.

6) (F) Formation of AI-powered weapons

The production of new weapons and automated offensive devices or automata generates several risks: the potential for world domination by the country that first acquires a novel technology offering higher power of destruction or other forms of dominance is only one. Other risks, such as a drift toward dystopic societies with ultra-authoritarian control by robotic units, have been considered by various science fiction works and movies. The risks' severity varies from medium to high, but the likelihood is not easy to assess. In this situation, we prefer to err on the side of caution, evaluating the odds as medium in the short and medium term and high in the long term.

7) (G) Growing a divide between integrated and anti-tech citizens

New technology brings profound consequences for our way of living, which are welcomed by some and repelled by others. The polarizing effect of new tools that significantly empower humans with new abilities and skills can have divisive effects when those tools are not available to everybody. Still, the same effect arises when they present features that pose the need for the user to decide for or against them. Factors that may create a division are the fear of government control or the potential harm to one's health.: we are familiar with these themes, as we have recently witnessed similar polarizations during the COVID-19 vaccination campaigns. The development of BCIs and under-skin microchips are examples of technologies that may generate a two-tiered society, with individuals who may reap its benefits and others who get marginalized by not having access to them. This may have long-term nocuous consequences on the general happiness, integration, and democracy of our future society. We may again classify this as a collateral downside of the growing pains toward a more empowered version of humans. However, the effect can be far-reaching and must concern us; in many cases, mitigating strategies exist, but they must be planned.

8) (H) Handover of control of critical systems to AI

The economic benefits of automation create a general tendency to substitute human decisions with algorithms. While the latter may show lower failure rates (as it has already happened in some cases, such as diagnosis of pathologies from clinical images) and therefore be overall beneficial in regular operation, here we are concerned with the fact that such a substitution, and the automation of decision procedures, involves risks associated with misbehavior or malfunction not considered in the design phase, and that produce catastrophic effects. One example of this kind is the famous failure of an early-warning system detecting the launch of ICBM missiles by the US toward the Soviet Union. On September 26, 1983, a rare alignment of the detection satellite with the sun and with the field of view of the US territory caused the system to report the launch of five missiles, when in fact, the detection was an artifact due to Sun glare[33]. Only the presence of a "human in the middle" (in the person of Colonel Stanislav Petrov) and his doubts about an attack involving only five missiles avoided a retaliation strike that would have likely caused the start of a global nuclear war. In general, the risk of overconfidence in automated decisions cannot be quantified if not connected with specific systems under control. However, we may still assess its potential severity as a medium overall. The likelihood of manifestation of nocuous effects can be estimated as medium in the short term and low/very low in longer time scales as we gain confidence in our validation systems and design more robust and reliable countermeasures.

9) (I) Inception of superintelligence with misaligned goals to those of humankind

This scenario has been discussed in some detail *supra* (Sec. IV and Sec. V). Developing a system with super-human capabilities in a broad enough range of tasks entails the possibility that the system acquires the skill to reflect on itself and its role in a wider context and develop objectives different from those it was initially designed to address. This might cause it to become hostile to humans or to pursue its goals in ways that conflict with the existence of human life on our planet. In both cases, this is an existential risk of the highest severity and, therefore, regardless of perceived or assessed likelihood, must be assigned very high values on our scale.

10) (J) Job automation effects on society

Several sources have discussed this global threat in detail; see, in particular, the books by Martin Ford [412], [413]. It is generally agreed that many tasks of repetitive nature –such as the operation and driving of vehicles for the transportation of humans or goods, the delivery of goods to the end users, intermediate processing tasks in production chains, and many others– are going to withstand significant change with the substitution of human operators with AI-driven systems, fueled by the largely associated decrease in cost; this process has already started, and it will only intensify in the short term. Similarly, the AI impact on education will be explicit in the widespread availability of real-time speech translation, the acquired higher trust of large language model outputs, and the automated production of educational audio/video products. Many experts argue that humans will not be substituted but rather empowered by AI assistance, which will have a beneficial and qualifying effect on their work conditions. Others note that AI will reduce the expertise required to perform complex tasks by human workers, depreciating the value of their experience and skills. It remains to be seen what the net

[33]https://en.wikipedia.org/wiki/1983_Soviet_nuclear_false_alarm_incident .

outcome of the ongoing transition will be. Still, if we only look at the likelihood and severity of adverse outcomes in this area, we must assess these as a medium in both cases. This is likely also to be a transient effect, so we believe the impact will decrease in the long term to a low-severity one.

B. DISCUSSION

The numerical evaluations of the risk of AI-related phenomena and their perceived likelihoods are manifestly subjective, and by themselves, each of them means very little. What matters is the big picture: by doing the exercise of asking ourselves the questions that must be answered to fill data into Table 1, we get to see how wide-ranging and far-reaching the potential consequences of the development of artificially intelligent systems are: we are led to consider that in every area of human activity, the introduction of AI may have serious drawbacks. We believe that by fostering an awareness of this fact and continuously monitoring the evolving status of each sub-area, we may successfully lay the basis for possible governance. As AI expert Stuart Russell aptly put it in his book "Human Compatible: AI and the Problem of Control" [403], recently, "everybody and their uncle" have been proposing supervisory organizations meant to oversee and control the development and the integration of AI in human activities. While he notes that those efforts were, in the beginning, ineffective due to their grass-root nature and their limited reach and impact, he admits in a post-scriptum that today, more organized and institutionalized entities are getting established, which have a chance to be able to influence AI developments and steer us more effectively away from unwanted, harmful situations. We certainly agree that such organizations are essential in our effort to mitigate or avert most potential dangers. Yet, we suggest that they may be insufficient to thwart the biggest ones, which may be out of reach of governance bodies. For example, we may consider the threat of AI-eased authoritarian involution: regimes will likely not collaborate with those organizations if they value the opportunity to get more robust control and power. Instead, they will likely feign their cooperation and diligence while secretly pursuing developments to reinforce control over society. A similar effect may happen with risk F in Table 1, the development of AI weapons. A likely result will be that countries that limit their technological developments by abiding by the rules set by intergovernmental organizations will have no means to gauge what happens worldwide and will be at a disadvantage compared to others.

Let us consider the light red and dark red boxes in Fig. 13, which correspond in our assessment to the biggest threats to humankind from AI development and deployment; not surprisingly, those threats are also the hardest to control and reduce. Also, not surprisingly, the risk that gets the highest score from our assessment is the development of a misaligned or malignant AGI; this is true regardless of what temporal horizon we consider. In other words, despite its very low likelihood, we are most concerned by that development as it is potentially the most harmful and destructive. This is due to its far-reaching consequences and the absence of possible countermeasures *ex-post*. We can only hope that the strategies that have started to be implemented today [6], and the considerate action of individuals at the driver's seat of companies who lead development efforts of AGI systems, will help reduce that risk.

A word must also be spent on three other risks that reach in our assessment a level of light or dark orange (respectively, level 7 and 8) in the short term (*i.e.*, within the next ten years). These are risk A (Acceleration of nocuous impact on the environment), risk F (Formation of AI weapons), and J (Job automation effects on society).

As far as the AI-development-induced nocuous impact on the environment is concerned, this is an effect that compounds the dire situation we are already facing in the world today, with soaring CO_2 emissions in the atmosphere, deforestation, and pollution of the environment. We believe that this risk has already become a reality. Although it is, in principle, never too late to fight the current situation, *e.g.*, by embracing the idea of de-growth [414], [415]—a controlled downscaling of production and consumption in the wealthiest countries—, we observe that the 180-degree turn in the way the world's leading economies work that would be required is practically impossible to achieve.

The development and accumulation of AI-powered weapons is also a danger that is already clear and present. As discussed in Sec. V, weapons development already includes AI among its technological ingredients. In connection to risk C (the use of AI for creating ultra-authoritarian regimes), this is a significant threat that is hard to mitigate because of the lack of instruments and the opacity of the involved processes. At the time of writing, given the very uncertain geopolitical situation we are currently facing, it is one of our most serious concerns.

Finally, concerning the risk of social disruptions due to the automation of an increasing number of jobs, this is a situation that rich countries can indeed effectively act upon by foreseeing social amortizers, universal income, and other measures. It is a problem that can be solved by throwing money at it, but this requires careful planning and a favorable political situation to be implemented. So here is an AI-related threat to our society that is very concrete, quite likely to manifest itself in the short term, and one we can defuse if we make it more manifest and if we clarify what options are possible –the kind of objectives we are trying to aim at with the present text. On the other hand, there is ground to be pessimistic in this case, as there have already been situations where job automation caused significant disruption, *e.g.*, at Amazon warehouses or McDonald's restaurants [413] –job loss, devaluation of human skills, worsening of working conditions– without government interventions to mitigate those effects.

1) An optimistic vision on the incoming revolution

We believe it is appropriate to conclude this long review on a positive note by echoing some of the points in the

previous sections about the enormous benefits that artificial intelligence systems and tools have brought to our modern society. Indeed, AI has profoundly revolutionized our lives within a mere decade, and the process is accelerating. It has significantly increased the living standards of human beings worldwide and across almost all income categories (although, unfortunately, this has come at least in part at the expense of most other animal species on this planet). For example, it is estimated that 85% of the world population today owns a smartphone, and an even more significant fraction can freely access a computer connected to the web. This provides humans with powerful AI-powered functionalities, such as finding one's way regardless of where one is on the planet, instantly translating text, generating computer code, constructing images or videos, solving problems with the help of LLMs, and learning from a vast database of documents and videos. Compiling a list of the benefits we have been reaping from AI technology indeed feels silly, no less than cavemen extolling the virtues of fire.

Within the boundaries of scientific research, some analysis of the present and foreseen future benefits of AI systems has been offered throughout Sec. IV. To summarize that discussion here, we note that in the scientific research areas we discussed in this document, as well as in others we did not comment on, the advantages brought by deeper analysis, more accessible and faster achievement of results, improved performance, higher precision that AI methods can provide over previous standards are so vast that it is hard even to start putting something on the other arm of the scale. This does not even consider all those situations where AI-powered technologies have generated genuine breakthroughs and paradigm-changing advancements that defy the definition of a pre-AI comparison point. We mention the following three:

1) AlphaFold and the Protein Folding Prediction, announced in December 2020 by DeepMind [416], is probably the most glaring example. Protein folding is a critical problem in biology, as proteins' enormously complex three-dimensional structure largely determines their function. AlphaFold demonstrated high precision in predicting the structure of specific protein structures, vastly outperforming traditional methods. This advancement bears profound implications for drug discovery, understanding diseases, and designing novel therapies.
2) Image recognition and computer vision have been thoroughly revolutionized by deep learning methods. Convolutional neural networks have achieved unprecedented accuracy in tasks ranging from object recognition to image classification and segmentation. These advancements have brought a wealth of benefits, *e.g.*, to clinical medicine, improving the precision of the diagnosis of patients, satellite imagery analysis, space exploration, autonomous driving, and pure research in several areas.
3) NLP tools and large language models such as the series of OpenAI's generative pre-trained transformers have wholly transformed language processing tasks. They found application to a wide area of research tasks in many different fields. For example, in biomedical research, NLP techniques have been applied to the analysis of large volumes of biomedical literature, electronic health records, and clinical notes, extracting valuable information from unstructured text, enabling researchers to identify disease patterns, predict outcomes, and improve clinical decision [417], [418].

The above examples should suffice to make the point that in scientific research, the development of AI technologies has produced massively positive developments across the board, and it is likely to continue to do so at an increasing pace. Suppose we broaden our field of view to consider the impact of those research enhancements on our society. In that case, we see distinctly positive outcomes already at the consumer end, notably those connected to better, more effective medicine. In the future, we can only expect a broader impact, with a speed-up of several procedures and bottlenecks of current research and development in all areas of scientific investigation carrying over their effect to an improved quality of human life.

Human progress is hard to define from within the time scale of the human lifetime; its precise appraisal also requires the vantage point of *ex-post* considerations. Yet we entertain no doubts that artificial intelligence is a necessary, unavoidable evolutionary step in the march of humankind toward its future. It is a future that may await us with dystopian or idyllic features. We thus feel we are today at a critical juncture when that future can still be –at least in part– shaped by our decisions and policies. As the introduction notes, shaping may have unimaginable and far-reaching long-term consequences in this area of the universe. For those reasons, artificial intelligence and its development must be kept on the discussion table in all human occupations and at all levels for the years to come. As scientists, we feel we have the responsibility to inform those discussions and keep them rational and pragmatic. Still, we also need to ensure a place in the decision-making procedures to defend science-driven reasons and their impact on those decisions from the attack of irrational arguments. We are therefore happy to see that several recent spontaneous initiatives and organizations worldwide are starting to work in that direction together with governments and inter-governmental bodies. This document is our small contribution to inform those discussions.

2) Limitations and potential biases of this study

As a position paper, this work may not be as unbiased and objective as a typical research work. It represents the viewpoint of a group of experienced scholars who have collectively published numerous scientific works across various disciplines over the past forty years. We are expressing our perspectives on the impacts that artificial intelligence is having on our research fields. It is important to note that our backgrounds

may influence our viewpoints. However, our collective views could be valuable and interesting to the community. This aspect is both a limitation and a valuable asset of our work.

ACKNOWLEDGMENT

A. Ustyuzhanin was supported by the Ministry of Education, Singapore, under its funding for the Research Centre of Excellence Institute for Functional Intelligence Materials and by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2022-028).

C. Casonato's research is supported in part by the activities of the MUR PNRR project FAIR - Future AI Research (PE00000013) funded by the Next Generation EU.

G. Olivato's contribution is realized with co-financing from the European Union – FSE REACT- EU, PON Ricerca e Innovazione 2014-2020.

Pietro Vischia's work was supported by the "Ramón y Cajal" program under Project No. RYC2021-033305-I funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

The research of J.J. Nieto was supported by the Agencia Estatal de Investigacion (AEI) of Spain Grant PID2020-113275GB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", by the "European Union" and Xunta de Galicia, grant ED431C 2023/12 for Competitive Reference Research Groups (2023–2026).

The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

REFERENCES

[1] D. C. Hsu, E. B. Ford, D. Ragozzine, and K. Ashby, "Occurrence rates of planets orbiting FGK stars: Combining Kepler DR25, Gaia DR2, and Bayesian Inference," *The Astronomical Journal*, vol. 158, no. 3, pp. 109–128, 2019.

[2] G. Dalrymple, "The age of the Earth in the twentieth century: a problem (mostly) solved," *Special Publications, Geological Society of London*, vol. 190, no. 1, pp. 205–221, 2001.

[3] M. S. Dodd, D. Papineau, T. Grenne *et al.*, "Evidence for early life in Earth's oldest hydrothermal vent precipitates," *Nature*, vol. 543, no. 7643, pp. 60–64, 2017.

[4] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[5] M. Tegmark, *Life 3.0 – Being Human in the Age of Artificial Intelligence*. Knopf, 2017.

[6] Council of the European Union, "REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down harmonised rules on artificial intelligence and amending regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)," *PE-CONS 24/24, 2021/0106(COD)*, 2024.

[7] J. Khalffa, *What is Intelligence?* Cambridge: White Lotus Press, 1994.

[8] S. Legg and M. Hudder, "Universal intelligence: a definition of machine intelligence," *Minds & Machines*, vol. 17, pp. 391–444, 2007.

[9] E. Ermer, L. Cosmides, and J. Tooby, "Functional specialization and the adaptationist program," in *The evolution of mind: Fundamental questions and controversies*, S. W. Gangestad and J. A. Simpson, Eds. The Guilford Press, 2017, pp. 153–160.

[10] G. Vallortigara, *Born Knowing*. Cambridge, MA: MIT Press, 2021.

[11] E. Versace, A. Martinho-Truswel, A. Kacelnik, and G. Vallortigara, "Priors in animal and artificial intelligence: Where does learning begin?" *Trends in Cognitive Sciences*, vol. 22, pp. 963–965, 2018.

[12] N. J. Mackintosh, "Intelligence in evolution," in *What is intelligence?*, J. Khalffa, Ed. Cambridge University Press, 1994.

[13] D. S. Wilson *et al.*, "Multilevel cultural evolution: From new theory to practical applications," *Proceedings of the National Academy of Science*, vol. 120, no. 16, p. e2218222120, 2023.

[14] P. Wang, "On defining artificial intelligence," *Journal of Artificial General Intelligence*, vol. 10, no. 2, pp. 1–37, 2019.

[15] OECD, "Explanatory memorandum on the updated OECd definition of an AI system," OECD Artificial Intelligence Papers, 8, OECD Publishing, Paris, 2024.

[16] C. Casonato, "Costituzione e intelligenza artificiale: Un'agenda per il prossimo futuro," *BioLaw Journal - Rivista Di BioDiritto*, vol. 2S, pp. 711–725, 2019.

[17] L. Floridi, "AI as agency without intelligence: on ChatGPT, large language models and other generative models," *Philosophy and Technology*, vol. 36, p. 15, 2023.

[18] N. Block, "How many concepts of consciousness?" *Behavioral and Brain Sciences*, vol. 18, no. 2, pp. 272–284, 1995.

[19] T. Nagel, *What is it like to be a bat? Mortal Questions*. Cambridge University Press, 1991.

[20] D. Chalmers, "Facing up to the problem of consciousness," *Journal of Consciousness Studies*, vol. 2, pp. 200–219, 1995.

[21] S. C. Hayes and S. G. Hofmann, "A biphasic relational approach to the evolution of human consciousness," *International Journal of Clinical and Health Psychology*, vol. 24, no. 4, p. 100380, 2023.

[22] G. Vallortigara, "Sentience does not require "higher" cognition. Commentary on Marino on Thinking Chickens," *Animal Sentience*, vol. 30, p. 6, 2017.

[23] ——, "Lessons from miniature brains: cognition cheap, memory expensive (sentience linked to active movement?)," *Animal Sentience*, vol. 29, p. 17, 2020.

[24] L. Weiskrantz, *Consciousness Lost and Found: A Neuropsychological Exploration*. Oxford: Oxford University Press, 1999.

[25] T. M. Schubert, D. Rothlein, T. Brothers, E. L. Coderre, K. Ledoux, B. Gordon, and M. McCloskey, "Lack of awareness despite complex visual processing: evidence from event-related potentials in a case of selective metamorphopsia," *Proceedings of the National Academy of Sciences of USA*, vol. 117, pp. 16 055–16 064, 2020.

[26] B. Merker, "Consciousness without a cerebral cortex: a challenge for neuroscience and medicine," *Behavioral Brain Sciences*, vol. 30, pp. 63–134, 2007.

[27] B. J. Baars, *A Cognitive Theory of Consciousness*. Cambridge University Press, 1993.

[28] G. Tononi, M. Boly, M. Massimini, and C. Koch, "Integrated information theory: from consciousness to its physical substrate," *Nature Reviews Neuroscience*, vol. 17, pp. 450–461, 2016.

[29] S. Dehaene and L. Naccache, "Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework," *Cognition*, vol. 79, pp. 1–37, 2001.

[30] A. B. Barrett and P. A. Mediano, "The Phi measure of integrated information is not well-defined for general physical systems," *Journal of Consciousness Studies*, vol. 26, pp. 11–20, 2019.

[31] B. Merker, K. Williford, and D. Rudrauf, "The integrated information theory of consciousness: a case of mistaken identity," *Behavioral Brain Sciences*, vol. 45, no. e41, 2022.

[32] M. Baron and M. Devor, "Might pain be experienced in the brainstem rather than in the cerebral cortex?" *Behavioural Brain Research*, vol. 427, p. 113861, 2022.

[33] A. Damasio, *Self Comes to Mind*. New York: Pantheon, 2010.

[34] A. Damasio and G. B. Carvalho, "The nature of feelings: evolutionary and neurobiological origins," *Nature Reviews Neuroscience*, vol. 14, pp. 143–152, 2013.

[35] A. Damasio, T. J. Grabowski, A. Bechara, H. Damasio, L. L. B. Ponto, J. Parvizi, and R. D. Hichwa, "Subcortical and cortical brain activity during the feeling of self-generated emotions," *Nature Neuroscience*, vol. 3, pp. 1049–1056, 2000.

[36] A. Damasio, H. Damasio, and D. Tranel, "Persistence of feelings and sentience after bilateral damage of the insula," *Cerebral Cortex*, vol. 23, no. 4, pp. 833–846, 2012.

[37] M. Solms, *The Hidden Spring*. London: Profile Book Ltd, 2021.

[38] E. von Holst and H. Mittelstaedt, "Das reafferenzprincip (wechselwirkungen zwischen zentralnervensystem und peripherie)," *Naturwissenschaften*, vol. 37, pp. 464–476, 1950.
[39] G. Vallortigara, "The rose and the fly. a conjecture on the origin of consciousness," *Biochemical and Biophysical Research Communications*, vol. 564, pp. 170–174, 2021.
[40] ——, "The efference copy signal as a key mechanism for consciousness," *Frontiers in Systems Neuroscience*, vol. 26, 2021.
[41] ——, *The Origins of Consciousness. Thoughts of the Crooked-Headed Fly.* London: Routledge, 2024.
[42] N. Humphrey, *Sentience. The invention of consciousness*. Cambridge, MA, USA: MIT Press, 2023.
[43] G. Vallortigara, "Visual cognition and representation in birds and primates," in *Vertebrate Comparative Cognition: Are Primates Superior to Non-Primates?*, L. J. Rogers and G. Kaplan, Eds. New York: Kluwer Academic/Plenum Publishers, 2004, pp. 57–94.
[44] O. Güntürkün and T. Bugnyar, "Cognition without cortex," *Trends Cognitive Sciences*, vol. 20, pp. 291–303, 2016.
[45] A. Nieder, L. Wagener, and P. Rinnert, "A neural correlate of sensory consciousness in a corvid bird," *Science*, vol. 369, pp. 1626–1629, 2020.
[46] A. S. Reber, *The First Minds: Caterpillars, 'Karyotes, and Consciousness*. New York: Oxford University Press, 2019.
[47] A. B. Barron and C. Klein, "What insects can tell us about the origins of consciousness," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 18, pp. 4900–4908, 2016.
[48] T. Nagel, "What is it like to be a bat?" *The Philosophical Review*, vol. 83, pp. 435–450, 1974.
[49] S. G. Hofmann and S. N. Doan, *The Social Foundations of Emotion: Developmental, Cultural, and Clinical Dimensions*. Washington, D.C.: American Psychological Association, 2018.
[50] G. A. Mashour, P. Roelfsema, J. P. Changeux, and S. Dehaene, "Conscious processing and the global neuronal workspace hypothesis," *Neuron*, vol. 105, pp. 776–798, 2020.
[51] S. Dehaene, C. Sergent, and J. P. Changeux, "A neuronal network model linking subjective reports and objective physiological data during conscious perception," *Proceedings of the National Academy of Sciences USA*, vol. 100, pp. 8520–8525, 2003.
[52] G. Tononi, "Consciousness as integrated information: a provisional manifesto," *Biological Bulletin*, vol. 215, pp. 216–242, 2008.
[53] ——, "Integrated information theory of consciousness: an updated account," *Archives Italiennes de Biologie*, vol. 150, pp. 293–329, 2012.
[54] S. Dehaene, H. Lau, and S. Kouider, "What is consciousness, and could machines have it?" *Science*, vol. 358, no. 6362, pp. 486–492, 2017.
[55] B. Odegaard, M. Y. Chang, H. Lau, and S. H. Cheung, "Inflation versus filling-in: why we feel we see more than we actually do in peripheral vision," *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 2018.
[56] J. E. LeDoux and R. Brown, "A higher-order theory of emotional consciousness," *Proceedings of the National Academy of Sciences USA*, vol. 114, pp. E2016–E2025, 2017.
[57] J. Morrison, "Perceptual confidence," *Analytic Philosophy*, vol. 78, pp. 99–147, 2016.
[58] H. Lau and D. Rosenthal, "Empirical support for higher-order theories of conscious awareness," *Trends in Cognitive Science*, vol. 15, pp. 365–373, 2011.
[59] A. Seth and T. Bayne, "Theories of consciousness," *Nature Review Neuroscience*, vol. 23, pp. 439–452, 2022.
[60] L. Polansky, W. Kilian, and G. Wittemyer, "Elucidating the significance of spatial memory on movement decisions by african savannah elephants using state-space models," *Proceedings of the Royal Society B*, vol. 282, p. 20143042, 2015.
[61] F. X. Neubert, R. B. Mars, A. G. Thomas, J. Sallet, and M. F. Rushworth, "Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex," *Neuron*, vol. 81, pp. 700–713, 2014.
[62] G. N. Elston, "Cortex, cognition and the cell: new insights into the pyramidal neuron and prefrontal function," *Cerebral Cortex*, vol. 13, pp. 1124–1138, 2003.
[63] W. von Humboldt, *Über die Verschiedenheit des menschlichen Sprachbaus und ihren Einfluss auf die geistige Entwicklung des Menschengeschlechts*. Bonn: Dümmlers Verlag, 1837, reprinted in 1960.
[64] L. Wittgenstein, *Philosophische Untersuchung*. Oxford, UK: Basil Blackwell, 1953.
[65] N. Chomsky, *Cartesian Linguistics: a Chapter in the History of Rationalist Thought*. New York: Harper and Row Pub. Inc, 1966.
[66] J. Fodor, *The Language of Thought*. Cambridge, MA: Harvard University Press, 1975.
[67] S. Pinker, *The Language Instinct. How the Mind Creates Language*. New York: William Morrow, 1994.
[68] A. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, pp. 433–460, 1950.
[69] W. J. Hutchins, *Machine Translation: Past, Present, and Future*. New York: Chichester, 1986.
[70] N. Chomsky, *Syntactic Structures*. The Hague, NL: Mouton & Co., 1957.
[71] ——, *Current Issues in Linguistic Theory*. The Hague, NL: Mouton & Co., 1964.
[72] T. Kasami, "An efficient recognition and syntax algorithm for context-free languages," AFCRL, Bedford, MA, Tech. Rep. 65-758, 1965.
[73] D. H. Younger, "Recognition and parsing of context-free languages in time $o(n^3)$," *Information and Control*, vol. 10, pp. 447–474, 1967.
[74] J. Earley, "An efficient context-free parsing algorithm," *Communications of the Association for Computing Machinery*, vol. 13, pp. 94–102, 1970.
[75] T. Winograd, *Understanding Natural Language*. New York: Academic Press, 1972.
[76] H. Alshawi, "Resolving quasi logical forms," *Computational Linguistics*, vol. 16, pp. 133–144, 1990.
[77] S. Lappin and H. J. Leass, "An algorithm for pronominal anaphora resolution," *Computational Linguistics*, vol. 20, pp. 535–561, 1994.
[78] S. Banerjee and T. Pedersen, "Extended gloss overlaps as a measure of semantic relatedness," in *Proceedings of the 18th international joint conference on Artificial intelligence*, 2003, pp. 805–810.
[79] A. Lascarides and N. Asher, "Segmented discourse representation theory: Dynamic semantics with discourse structure," in *Computing Meaning*, H. Bunt and R. Muskens, Eds. Berlin: Springer-Verlag, 2007, pp. 87–124.
[80] D. E. Rumelhart and J. L. McClelland, "On learning the past tenses of english verbs," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press, 1986, vol. 2, pp. 216–271.
[81] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, pp. 179–221, 1990.
[82] R. Miikkulainen and M. G. Dyer, "Natural language processing with modular PDP networks and distributed lexicon," *Cognitive Science*, vol. 15, pp. 343–399, 1991.
[83] J. L. Elman, "Grammatical structure and distributed representations," in *Connectionism: Theory and Practice*, S. Davis, Ed. Oxford, UK: Oxford University Press, 1992.
[84] K. Plunkett and V. A. Marchman, "From rote learning to system building: Acquiring verb morphology in children and connectionist nets," *Cognition*, vol. 48, pp. 21–69, 1993.
[85] B. MacWhinney, "The dinosaurs and the ring," in *The Reality of Linguistics Rules*, R. Corrigan, G. Iverson, and S. Lima, Eds. Amsterdam: John Benjamins, 1994, pp. 283–320.
[86] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.
[87] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2014, pp. 1724–1734.
[88] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proceedings of the International Speech Communication Association Conference*, 2013, pp. 2345–2349.
[89] G. Saon *et al.*, "English conversational telephone speech recognition by humans and machines," in *Proceedings of the International Speech Communication Association Conference*, 2017, pp. 132–136.
[90] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.
[91] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.
[92] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations*, 2016.

[93] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan *et al.*, "Language models are few-shot learners," *arXiv*, vol. abs/2005.14165, 2020.

[94] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," in *Proceedings of Advances in Neural Information Processing Systems*, 2022, pp. 27 730–27 744.

[95] D. Jannai, A. Meron, B. Lenz, Y. Levine, and Y. Shoham, "Human or not? a gamified approach to the Turing test," 2023. [Online]. Available: https://arxiv.org/abs/2305.20010

[96] E. M. Bender and A. Koller, "Climbing towards NLU: On meaning, form, and understanding in the age of data," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Somerset, NJ: Association for Computational Linguistics, 2020, pp. 5185–5198.

[97] A. Søgaard, "Understanding models understanding language," *Synthese*, vol. 200, 2022, Art. no. 443.

[98] M. Tamir and E. Shech, "Machine understanding and deep learning representation," *Synthese*, vol. 201, p. 51, 2023.

[99] A. Søgaard, "Grounding the vector space of an octopus: Word meaning from raw text," *Minds and Machines*, vol. 33, pp. 33–54, 2023.

[100] A. Srivastava *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," 2023. [Online]. Available: https://arxiv.org/abs/2206.04615

[101] T. Hagendorff, I. Dasgupta, M. Binz, S. C. Y. Chan, A. Lampinen, J. X. Wang, Z. Akata, and E. Schulz, "Machine psychology," 2024. [Online]. Available: https://arxiv.org/abs/2303.13988

[102] R. Bommasani, P. Liang, and T. Lee, "Holistic evaluation of language models," *Annals of the New York Academy of Sciences*, pp. 1–7, 2023.

[103] G. W. Lindsay and D. Bau, "Testing methods of neural systems understanding," *Cognitive Systems Research*, vol. 82, 2023, Art. no. 101156.

[104] M. Kosinski, "Evaluating large language models in Theory of Mind tasks," 2024. [Online]. Available: https://arxiv.org/abs/2302.02083

[105] A. Marchetti, C. Di Dio, A. Cangelosi, F. Manzi, and D. Massaro, "Developing ChatGPT's theory of mind," *Frontiers in Robotics and AI*, vol. 10, p. 1189525, 2023.

[106] S. Trott, C. Jones, T. Chang, J. Michaelov, and B. Bergen, "Do large language models know what humans know?" *Cognitive Science*, vol. 47, p. e13309, 2023.

[107] D. Chalmers, "Could a Large Language Model be conscious?" 2024. [Online]. Available: https://arxiv.org/abs/2303.07103

[108] J. Aru, M. E. Larkum, and J. M. Shine, "The feasibility of artificial consciousness through the lens of neuroscience," *Trends in Neuroscience*, vol. 46, pp. 1008–1017, 2023.

[109] R. VanRullen, "Perception science in the age of deep neural networks," *Frontiers in Psychology*, vol. 8, 2017, DOI 10.3389/fpsyg.2017.00142.

[110] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of Advances in Neural Information Processing Systems*, 2012, pp. 1090–1098.

[111] A. Andreopoulos and J. K. Tsotsos, "50 years of object recognition: Directions forward," *Computer Vision and Image Understanding*, vol. 117, pp. 827–891, 2013.

[112] S. M. Kosslyn, *Image and Mind*. Cambridge, MA: Harvard University Press, 1980.

[113] Z. Pylyshyn, "What the mind's eye tells the mind's brain: A critique of mental imagery," *Psychological Bulletin*, vol. 80, 1973, DOI 10.1037/h0034650.

[114] S. M. Kosslyn, *Image and Brain: the Resolution of the Imagery Debate*. Cambridge, MA: MIT Press, 1994.

[115] S. T. Moulton and S. M. Kosslyn, "Imagining predictions: mental imagery as mental emulation," *Philosophical Transactions of the Royal Society B*, vol. 364, pp. 1273–1280, 2009.

[116] J. Pearson and S. M. Kosslyn, "The heterogeneity of mental representation: Ending the imagery debate," *Proceedings of the National Academy of Sciences USA*, vol. 112, pp. 10 089–10 092, 2015.

[117] N. Dijkstra, S. E. Bosch, and M. A. van Gerven, "Shared neural mechanisms of visual perception and imagery," *Trends in Cognitive Sciences*, vol. 23, pp. 423–434, 2019.

[118] S. Hurley, "The shared circuits model (scm): how control, mirroring, and simulation can enable imitation, deliberation, and mindreading," *Behavioral and Brain Sciences*, vol. 31, no. 1, pp. 1–22, 2008.

[119] G. Hinton and R. S. Zemel, "Autoencoders, minimum description length and Helmholtz free energy," in *Proceedings of Advances in Neural Information Processing Systems*, 1994, pp. 3–10.

[120] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 28, pp. 504–507, 2006.

[121] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[122] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, pp. 1–127, 2009.

[123] K. Fukushima, "Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, pp. 193–202, 1980.

[124] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2010, pp. 7–15.

[125] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high-level feature learning," in *Proceedings of the International Conference on Computer Vision*, 2011, pp. 6–14.

[126] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 818–833.

[127] A. Plebe and M. Da Lio, "On the road with 16 neurons: Towards interpretable and manipulable latent representations for visual predictions in driving scenarios," *IEEE Access*, vol. 8, pp. 179 716–179 734, 2020.

[128] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Berlin: Springer-Verlag, 2015, pp. 234–241.

[129] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. B. Tenenbaum, "Deep convolutional inverse graphics network," in *Proceedings of Advances in Neural Information Processing Systems*, 2015, pp. 2539–2547.

[130] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *Proceedings of the International Conference on Learning Representations*, 2014.

[131] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of Machine Learning Research*, E. P. Xing and T. Jebara, Eds., 2014, pp. 1278–1286.

[132] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," in *Proceedings of Advances in Neural Information Processing Systems*, 2014.

[133] I. Gulrajani, K. Kumar, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "PixelVAE: A latent variable model for natural images," in *Proceedings of the International Conference on Learning Representations*, 2018.

[134] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the International Conference on Machine Learning*, 2015.

[135] C. Jarzynski, "Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach," *Physical Review E*, vol. 56, pp. 5018–5035, 1997.

[136] A. Bansal, E. Borgnia, H.-M. Chu, J. S. Li, H. Kazemi, F. Huang, M. Goldblum, J. Geiping, and T. Goldstein, "Cold Diffusion: Inverting arbitrary image transforms without noise," 2022. [Online]. Available: https://arxiv.org/abs/2208.09392

[137] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proceedings of Advances in Neural Information Processing Systems*, 2020.

[138] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 684–10 695.

[139] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proceedings of Advances in Neural Information Processing Systems*, 2021, pp. 8780–8794.

[140] A. Ramesh *et al.*, "Zero-shot text-to-image generation," in *Proceedings of Machine Learning Research*, 2021, pp. 8821–8831.

[141] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations*, 2021.

[142] C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," in *Proceedings of Advances in Neural Information Processing Systems*, 2022, pp. 36 479–36 494.

[143] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660.

[144] A. Chatterjee, "Art in an age of artificial intelligence," *Frontiers in Psychology*, vol. 13, 2022, Art. no. 1024449.

[145] F. Tao, "A new harmonisation of art and technology: Philosophic interpretations of artificial intelligence art," *Critical Arts*, vol. 36, pp. 110–125, 2022.

[146] I. Kalpokas, "Work of art in the age of its AI reproduction," *Philosophy and Social Criticism*, pp. 1–19, 2023.

[147] J. Zylinska, "Art in the age of artificial intelligence – there is more to art than AI-created artifacts, but computational creativity is worth pursuing," *Science*, vol. 381, pp. 139–140, 2023.

[148] C. Casonato, "Unlocking the Synergy: Artificial Intelligence and (old and new) human rights," *BioLaw Journal - Rivista Di BioDiritto*, vol. 3, pp. 233–240, 2023.

[149] A. Plebe, H. Svensson, S. Mahmoud, and M. Da Lio, "Human-inspired autonomous driving: A survey," *Cognitive Systems Research*, vol. 83, 2024, Art. no. 101169. [Online]. Available: https://doi.org/10.1016/j.cogsys.2023.101169

[150] S. Pendleton *et al.*, "Perception, planning, control, and coordination for autonomous vehicles," *Machines*, vol. 5, no. 1, 2017, Art. no. 6.

[151] A. Marshall and A. Davies, "Uber's self-driving car didn't know pedestrians could jaywalk," *Wired*, 2019. [Online]. Available: https://www.wired.com/story/ubers-self-driving-car-didnt-know-pedestrians-could-jaywalk/

[152] US National Transport Safety Board, "Collision between vehicle controlled by developmental automated driving system and pedestrian," 2019. [Online]. Available: https://www.ntsb.gov/news/events/Pages/2019-HWY18MH010-BMG.aspx

[153] Z. Halim, R. Kalsoom, S. Bashir, and G. Abbas, "Artificial intelligence techniques for driving safety and vehicle crash prediction," *Artificial Intelligence Review*, vol. 46, pp. 351–372, 2016.

[154] K. Kreutz and J. Eggert, "Analysis of the generalized intelligent driver model (GIDM) for merging situations," in *Proceedings of the 2021 IEEE Intelligent Vehicles Symposium*, 2021, pp. 34–41.

[155] C. Menéndez-Romero, "Maneuver planning for highly automated vehicles," Ph.D. dissertation, Albert-Ludwigs-Universität Freiburg, 2021. [Online]. Available: https://freidok.uni-freiburg.de/data/222337

[156] E. Adams, "Why we're still years away from having self-driving cars," Vox, Sep. 25, 2020. [Online]. Available: https://www.vox.com/recode/2020/9/25/21456421/why-self-driving-cars-autonomous-still-years-away

[157] L. Eliot, "Whether those endless edge or corner cases are the long-tail doom for AI self-driving cars," Forbes, Jul. 13, 2021. [Online]. Available: https://www.forbes.com/sites/lanceeliot/2021/07/13/whether-those-endless-edge-or-corner-cases-are-the-long-tail-doom-for-ai-self-driving-cars/

[158] The New York Times, "Silicon valley is resetting expectations for self-driving cars and settling in for years of more work," May 25, 2021. [Online]. Available: https://www.nytimes.com/2021/05/25/business/silicon-valley-is-resetting-expectations-for-self-driving-cars-and-settling-in-for-years-of-more-work.html

[159] ——, "The costly pursuit of self-driving cars continues on and on and on," May 24, 2021. [Online]. Available: https://www.nytimes.com/2021/05/24/technology/self-driving-cars-wait.html

[160] A. M. Boggs, R. Arvin, and A. J. Khattak, "Exploring the who, what, when, where, and why of automated vehicle disengagements," *Accident Analysis and Prevention*, vol. 136, 2020, Art. no. 105406.

[161] M. Bojarski *et al.*, "End to End learning for self-driving cars," 2016. [Online]. Available: https://arxiv.org/abs/1604.07316

[162] M. Bansal and A. Ogale, "Learning to drive: Beyond pure imitation," 2018. [Online]. Available: https://medium.com/waymo/learning-to-drive-beyond-pure-imitation-465499f8bcb2

[163] R. Brooks, "A robust layered control system for a mobile robot," *IEEE Journal of Robotics and Automation*, vol. 2, no. 1, pp. 14–23, 1986.

[164] T. J. Prescott, P. Redgrave, and K. Gurney, "Layered control architectures in robots and vertebrates," *Adaptive Behavior*, vol. 7, no. 1, pp. 99–127, 1999.

[165] P. Cisek, "Cortical mechanisms of action selection: the affordance competition hypothesis," *Philosophical Transactions of the Royal Society B*, vol. 362, no. 1485, pp. 1585–1599, 2007.

[166] G. Pezzulo and P. Cisek, "Navigating the affordance landscape: Feedback control as a process model of behavior and cognition," *Trends in Cognitive Sciences*, vol. 20, no. 6, pp. 414–424, 2016.

[167] K. Meyer and A. Damasio, "Convergence and divergence in a neural architecture for recognition and memory," *Trends in Neurosciences*, vol. 32, no. 7, pp. 376–382, 2009.

[168] Dreams4Cars, "Dream-like simulation abilities for automated cars," 2020, accessed on 05/12/2024. [Online]. Available: https://www.dreams4cars

[169] M. Da Lio, R. Donà, G. P. Rosati Papini, and A. Plebe, "The biasing of action selection produces emergent human-robot interactions in autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1254–1261, 2022.

[170] M. Da Lio, A. Cherubini, G. P. Rosati Papini, and A. Plebe, "Complex self-driving behaviors emerging from affordance competition in layered control architectures," *Cognitive Systems Research*, vol. 79, pp. 4–14, 2023.

[171] R. Grush, "The emulation theory of representation: Motor control, imagery, perception," *Behavioral and Brain Sciences*, vol. 27, no. 3, pp. 377–396, 2004.

[172] G. Hesslow, "The current status of the simulation theory of cognition," *Brain Research*, vol. 1428, pp. 71–79, 2012.

[173] H. Svensson, S. Thill, and T. Ziemke, "Dreaming of electric sheep? exploring the functions of dream-like mechanisms in the development of mental imagery simulations," *Adaptive Behavior*, vol. 21, no. 4, pp. 222–238, 2013.

[174] G. P. Rosati Papini, A. Plebe, M. Da Lio, and R. Donà, "A reinforcement learning approach for enacting cautious behaviours in autonomous driving system: Safe speed choice in the interaction with distracted pedestrians," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–18, 2021.

[175] P. Wang, *Non-axiomatic logic: A model of intelligent reasoning*. World Scientific, 2013.

[176] ——, "A unified model of reasoning and learning," in *Proceedings of the International Workshop on Self-Supervised Learning*. PMLR, 2022, pp. 28–48.

[177] P. Wang and X. Li, "Different conceptions of learning: Function approximation vs. self-organization," in *Artificial General Intelligence: 9th International Conference, AGI 2016*. New York, USA: Springer International Publishing, 2016, pp. 140–149.

[178] J. R. Wolpaw and E. W. Wolpaw, Eds., *Brain Computer Interfaces: Principles and Practice*. Oxford University Press, 2012.

[179] C. S. Nam, A. Nijholt, and F. Lotte, Eds., *Brain–computer interfaces handbook: technological and theoretical advances*. CRC Press, 2018.

[180] M. Clerc, L. Bougrain, and F. Lotte, Eds., *Brain-computer interfaces 2: technology and applications*. John Wiley & Sons, 2016.

[181] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 4, no. 2, 2007, Art. no. R1.

[182] F. Lotte *et al.*, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10-year update," *Journal of Neural Engineering*, vol. 15, no. 3, 2018, Art. no. 031005.

[183] G. Pfurtscheller, D. Flotzinger, and J. Kalcher, "Brain-computer interface—a new communication device for handicapped persons," *Journal of Microcomputer Applications*, vol. 16, no. 3, pp. 293–301, 1993.

[184] Y. Roy *et al.*, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of Neural Engineering*, vol. 16, no. 5, 2019, Art. no. 051001.

[185] F. Yger, M. Berar, and F. Lotte, "Riemannian approaches in brain-computer interfaces: a review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1753–1762, 2016.

[186] M. Congedo, A. Barachant, and R. Bhatia, "Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review," *Brain-Computer Interfaces*, vol. 4, no. 3, pp. 155–174, 2017.

[187] R. Roy *et al.*, "Retrospective on the first passive brain-computer interface competition on cross-session workload estimation," *Frontiers in Neuroergonomics*, vol. 3, 2022, Art. no. 838342.

[188] C. Jeunet *et al.*, "A user-centred approach to unlock the potential of non-invasive bcis: an unprecedented international translational effort," in *CHIST-ERA Conference 2020*, 2020.

[189] X. Wei *et al.*, "2021 BEETL Competition: Advancing Transfer Learning for Subject Independence & Heterogenous EEG Data Sets," in *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 2022, pp. 205–219.

[190] G. Singh, R. N. Roy, and C. P. Chanel, "POMDP-based adaptive interaction through physiological computing," *Frontiers in Artificial Intelligence and Applications*, vol. 354, pp. 32–45, 2022.

[191] R. N. Roy, N. Drougard, T. Gateau, F. Dehais, and C. P. Chanel, "How can physiological computing benefit human-robot interaction?" *Robotics*, vol. 9, no. 4, 2020, Art. no. 100.

[192] B. F. Yuksel *et al.*, "Learn piano with bach: An adaptive learning interface that adjusts task difficulty based on brain state," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 5372–5384.

[193] C. Jeunet, B. N'Kaoua, R. N'Kambou, and F. Lotte, "Why and how to use intelligent tutoring systems to adapt MI-BCI training to each user," in *6th International BCI Meeting*, 2016.

[194] T. O. Zander, L. R. Krol, N. P. Birbaumer, and K. Gramann, "Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity," *Proceedings of the National Academy of Sciences*, vol. 113, no. 52, pp. 14 898–14 903, 2011.

[195] L. R. Krol, P. Haselager, and T. O. Zander, "Cognitive and affective probing: a tutorial and review of active learning for neuroadaptive technology," *Journal of Neural Engineering*, vol. 17, no. 1, 2020, Art. no. 012001.

[196] A. Roc *et al.*, "A review of user training methods in brain computer interfaces based on mental tasks," *Journal of Neural Engineering*, vol. 18, no. 1, 2021, Art. no. 011002.

[197] J. L. Collinger *et al.*, "High-performance neuroprosthetic control by an individual with tetraplegia," *The Lancet*, vol. 381, no. 9866, pp. 557–564, 2013.

[198] B. Wodlinger *et al.*, "Ten-dimensional anthropomorphic arm control in a human brain–machine interface: difficulties, solutions, and limitations," *Journal of Neural Engineering*, vol. 12, no. 1, 2014, Art. no. 016011.

[199] A. Benabid *et al.*, "An exoskeleton controlled by an epidural wireless brain–machine interface in a tetraplegic patient: a proof-of-concept demonstration," *The Lancet Neurology*, vol. 18, no. 12, pp. 1112–1122, 2019.

[200] A. B. Ajiboye *et al.*, "Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration," *The Lancet*, vol. 389, no. 10081, pp. 1821–1830, 2017.

[201] S. Flesher *et al.*, "A brain-computer interface that evokes tactile sensations improves robotic arm control," *Science*, vol. 372, no. 6544, pp. 831–836, 2021.

[202] F. R. Willett *et al.*, "High-performance brain-to-text communication via handwriting," *Nature*, vol. 593, no. 7858, pp. 249–254, 2021.

[203] ——, "A high-performance speech neuroprosthesis," *Nature*, vol. 620, pp. 1031–1036, 2023.

[204] S. L. Metzger *et al.*, "A high-performance neuroprosthesis for speech decoding and avatar control," *Nature*, vol. 620, pp. 1037–1046, 2023.

[205] J. D. R. Millán *et al.*, "Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges," *Frontiers in Neuroscience*, p. 161, 2010.

[206] R. Mane, T. Chouhan, and C. Guan, "Bci for stroke rehabilitation: motor and beyond," *Journal of Neural Engineering*, vol. 17, no. 4, 2020, Art. no. 041001.

[207] S. H. Fairclough and T. O. Zander, Eds., *Current Research in Neuroadaptive Technology*. Elsevier, 2021.

[208] A. Lécuyer *et al.*, "Brain-computer interfaces, virtual reality, and videogames," *Computer*, vol. 41, no. 10, pp. 66–72, 2008.

[209] B. Z. Allison and C. Neuper, "Could anyone use a bci?" in *Brain-Computer Interfaces: Applying our Minds to Human-Computer Interaction*, D. S. Tan and A. Nijholt, Eds. London: Springer, 2010, pp. 35–54.

[210] T. O. Zander and C. Kothe, "Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general," *Journal of Neural Engineering*, vol. 8, no. 2, 2011, Art. no. 025005.

[211] H. Cecotti, "Spelling with non-invasive Brain–Computer Interfaces–Current and future trends," *Journal of Physiology-Paris*, vol. 105, no. 1–3, pp. 106–114, 2011.

[212] J. Thielen, P. van den Broek, J. Farquhar, and P. Desain, "Broad-Band visually evoked potentials: re (con) volution in brain-computer interfacing," *PloS One*, vol. 10, no. 7, 2015, Art. no. e0133797.

[213] S. Nagel and M. Spüler, "World's fastest brain-computer interface: combining EEG2Code with deep learning," *PloS One*, vol. 14, no. 9, 2019, Art. no. e0221909.

[214] Z. Bai, K. N. Fong, J. J. Zhang, J. Chan, and K. H. Ting, "Immediate and long-term effects of BCI-based rehabilitation of the upper extremity after stroke: a systematic review and meta-analysis," *Journal of Neuroengineering and Rehabilitation*, vol. 17, pp. 1–20, 2020.

[215] F. Pichiorri *et al.*, "Brain–computer interface boosts motor imagery practice during stroke recovery," *Annals of Neurology*, vol. 77, no. 5, pp. 851–865, 2015.

[216] A. Biasiucci *et al.*, "Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke," *Nature Communications*, vol. 9, no. 1, 2018, Art. no. 2421.

[217] M. Musso *et al.*, "Aphasia recovery by language training using a brain–computer interface: A proof-of-concept study," *Brain Communications*, vol. 4, no. 1, 2022, Art. no. fcac008.

[218] L. George and A. Lécuyer, "An overview of research on "passive" brain-computer interfaces for implicit human-computer interaction," in *International Conference on Applied Bionics and Biomechanics, ICABB 2010 - Workshop W1 "Brain-Computer Interfacing and Virtual Reality"*, 2010.

[219] C. Mühl, B. Allison, A. Nijholt, and G. Chanel, "A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges," *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 66–87, 2014.

[220] R. Chavarriaga, A. Sobolewski, and J. D. R. Millán, "Errare machinale est: the use of error-related potentials in brain-machine interfaces," *Frontiers in Neuroscience*, 2014, Art. no. 208.

[221] A. Appriou *et al.*, "Towards measuring states of epistemic curiosity through electroencephalographic signals," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 4006–4011.

[222] E. Klein and A. Rubel, "Privacy and ethics in brain–computer interface research," in *Brain–Computer Interfaces Handbook: Technological and Theoretical Advances*, F. L. Chang S. Nam, Anton Nijholt, Ed. Boca Raton: CRC Press, 2018, pp. 653–668.

[223] T. Dorigo, *Anomaly! Collider Physics and the Quest for New Phenomena at Fermilab*. World Scientific, 2016.

[224] B. Denby, "Neural networks and cellular automata in experimental high energy physics," *Computer Physics Communications*, vol. 49, pp. 429–433, 1988.

[225] G. Moneti *et al.*, "Advanced analysis methods in particle physics," *Nucl. Phys. B (Proc. Suppl.)*, vol. 59, pp. 17–20, 1997.

[226] M. Acciarri *et al.*, "Search for the Standard Model Higgs boson in e+e- interactions at 161 <= sqrt(s) < 172 gev," *Physics Letters B*, vol. 411, no. 3-4, pp. 373–386, 1997.

[227] B. Denby, "Neural networks in high energy physics: a ten-year perspective," *Computer Physics Communications*, vol. 119, pp. 219–231, 1999.

[228] G. Aad and others (ATLAS Collaboration), "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," *Physics Letters B*, vol. 716, no. 1, pp. 1–29, 2012.

[229] S. Chatrchyan and others (CMS Collaboration), "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Physics Letters B*, vol. 716, no. 1, pp. 30–61, 2012.

[230] P. Calafiura, D. Rousseau, and K. Terao, Eds., *Artificial Intelligence for High-Energy Physics*. World Scientific, 2022.

[231] T. Dorigo *et al.*, "Toward the End-to-End Optimization of Particle Physics Instruments with Differentiable Programming," *Reviews in Physics*, vol. 10, 2023, Art. no. 100085.

[232] T. M. Wong *et al.*, "Ten to the power fourteen," IBM Research, Tech. Rep. RJ10502 (ALM1211-004), 2012.

[233] R. Kurtzweil, *The Singularity Is Near*. Viking, 2005.

[234] S. Agostinelli *et al.*, "Geant4 – a simulation toolkit," *Nucl. Instrum. Meth. A*, vol. 506, pp. 250–303, 2003.

[235] G. Kutyniok, "The Mathematics of Artificial Intelligence," 2022. [Online]. Available: https://arxiv.org/abs/2203.08890

[236] J. L. Wei *et al.*, "An optimal neural network design for fractional deep learning of logistic growth," *Neural Computing and Applications*, vol. 35, pp. 10 837–10 846, 2023.

[237] P. Grohs *et al.*, "A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of black-scholes partial differential equations," *Memoirs of the American Mathematical Society*, vol. 284, no. 1410, pp. 1–106, 2023.

[238] S. Ornes, "Peering inside the black box of AI," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 120, p. e2307432120, 2023.

[239] T. Wischmeyer, *Artificial Intelligence and Transparency: Opening the Black Box*. Springer, 2020.

[240] H. Couclelis, "Artificial intelligence in geography: Conjectures on the shape of things to come," *The Professional Geographer*, vol. 38, no. 1, pp. 1–11, 1986.

[241] J. P. Simon, "Artificial intelligence: scope, players, markets and geography," *Digital Policy, Regulation and Governance*, vol. 21, no. 3, pp. 208–237, 2019.
[242] Wikipedia, "Google Earth," n.d. [Online]. Available: https://en.wikipedia.org/wiki/Google_Earth
[243] B. Zhao, S. Zhang, C. Xh, Y. Sun, and C. Deng, "Deep fake geography? when geospatial data encounter Artificial Intelligence," *Cartography and Geographic Information Science*, vol. 48, no. 4, pp. 338–352, 2021.
[244] E. Felten, M. Raj, and R. Seamans, "Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses," *Strategic Management Journal*, vol. 42, no. 12, pp. 2195–2217, 2021.
[245] J. Arslan *et al.*, "Artificial intelligence algorithms for analysis of geographic atrophy: a review and evaluation," *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 57–57, 2020.
[246] M. C. Chen *et al.*, "Artificial intelligence and visual analytics in geographical space and cyberspace: Research opportunities and challenges," *Earth-Science Reviews*, vol. 241, 2023, Art. no. 104438.
[247] J. Rockström *et al.*, "Sustainable intensification of agriculture for human prosperity and global sustainability," *Ambio*, vol. 46, pp. 4–17, 2017.
[248] J. Sayer and K. G. Cassman, "Agricultural innovation to protect the environment," *Proceedings of the National Academy of Sciences USA*, vol. 110, pp. 8345–8348, 2013.
[249] B. Davis *et al.*, "Estimating global and country-level employment in agrifood systems," FAO Statistics Working Paper Series, Tech. Rep. 23-34, 2023.
[250] H. Barrett and D. C. Rose, "Perceptions of the fourth agricultural revolution: What's in, what's out, and what consequences are anticipated?" *Sociologia Ruralis*, vol. 62, pp. 162–189, 2020.
[251] G. Bannerjee, U. Sarkar, and S. Das, "Artificial intelligence in agriculture: A literature survey," *International Journal of Scientific Research in Computer Science Applications and Management Studies*, vol. 7, no. 3, 2018.
[252] V. Sachithra and L. D. Subhashini, "How artificial intelligence uses to achieve agricultural sustainability: Systematic review," *Artificial Intelligence in Agriculture*, vol. 8, pp. 46–59, 2023.
[253] M. T. Linaza *et al.*, "Data-driven artificial intelligence applications for sustainable precision agriculture," *Agronomy*, vol. 11, 2021, DOI 10.3390/agronomy11061227.
[254] D. E. Clay, S. Brugler, and B. Joshi, "Will artificial intelligence and machine learning change agriculture: A special issue," *Agronomy Journal*, vol. 116, pp. 791–794, 2024.
[255] S. Esposito, D. Carputo, T. Cardi, and P. Tripodi, "Applications and trends of machine learning in genomics and phenomics for next generation breeding," *Plants*, vol. 9, 2019, Art. no. 34.
[256] W. Yang *et al.*, "Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives," *Molecular Plant*, vol. 13, pp. 187–214, 2020.
[257] A. Tzachor *et al.*, "Responsible artificial intelligence in agriculture requires systemic understanding of risks and externalities," *Nature Machine Intelligence*, vol. 4, pp. 104–109, 2022.
[258] L. A. Puntel *et al.*, "How digital is agriculture in a subset of countries from South America? Adoption and limitations," *Crop & Pasture Science*, vol. 74, pp. 555–572, 2023.
[259] A. Tzachor *et al.*, "Large language models and agricultural extension services," *Nature Food*, vol. 4, pp. 941–948, 2023.
[260] K. Villiers *et al.*, "Advancing artificial intelligence to help feed the world," *Nature Biotechnology*, 2023.
[261] B. D. Hansen *et al.*, "Current status and future opportunities for digital agriculture in Australia," *Crop & Pasture Science*, vol. 74, pp. 524–537, 2023.
[262] V. Kaul, S. Enslin, and S. A. Gross, "History of artificial intelligence in medicine," *Gastrointestinal Endoscopy*, vol. 92, no. 4, pp. 807–812, 2020.
[263] E. A. Feigenbaum and B. G. Buchanan, "25th anniversary of the Dendral Project," The History of Artificial Intelligence - Spotlight at Stanford, n.d. [Online]. Available: https://exhibits.stanford.edu/ai/catalog/wr759qp9369
[264] SRI International, "Shakey the robot," 2024. [Online]. Available: https://www.sri.com/hoi/shakey-the-robot/
[265] W. van Melle, "MYCIN: A knowledge-based consultation program for infectious disease diagnosis," *International Journal of Man-Machine Studies*, vol. 10, no. 3, pp. 313–322, 1978.
[266] Institute of Medicine (US) Council on Health Care Technology, "Internist-1/Quick Medical Reference (QMR)," Medical Technology Assessment Directory: A Pilot Reference to Organizations, Assessments, and Information Resources, 1988. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK218496/
[267] G. L. Kuperman, B. B. Maack, K. Bauer, and R. M. Gardner, "The impact of the help computer system on the LDS Hospital Paper Medical Record," *Top Health Rec Manage*, vol. 12, no. 1, pp. 1–9, 1991. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/10112158/
[268] L. T. Powell, G. A. Diamond, P. K. Shah, and J. G. Ferguson, "Corsage: A critiquing system for coronary care," in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 1989. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245699/
[269] M. Savoy, "IDX-DR for diabetic retinopathy screening," *American Family Physician*, vol. 101, no. 5, pp. 307–308, 2020.
[270] A. Madden and A. Bekker, "Artificial Intelligence for EHR. Use Cases, Costs, Challenges," SCNSoft. [Online]. Available: https://www.scnsoft.com/healthcare/ehr/artificial-intelligence
[271] M. A. Al Antari, "Artificial intelligence for medical diagnostics—existing and future AI technology," *Diagnostics*, vol. 13, no. 4, 2023, DOI 10.3390/diagnostics13040688.
[272] Z. Jie, Z. Zhiying, and L. Li, "A meta-analysis of Watson for oncology in clinical application," *Scientific Reports*, vol. 11, 2021, Art. no. 5792.
[273] K. Theofilatos *et al.*, "Predicting protein complexes from weighted protein–protein interaction graphs with a novel unsupervised methodology: Evolutionary enhanced Markov clustering," *Artificial Intelligence in Medicine*, vol. 63, no. 3, pp. 181–189, 2015.
[274] T. Rapakoulia, "EnsembleGASVR: a novel ensemble method for classifying missense single nucleotide polymorphisms," *Bioinformatics*, vol. 30, no. 16, pp. 2324–2333, 2014.
[275] D. D. Luxton, "Recommendations for the ethical use and design of artificial intelligent care providers," *Artificial Intelligence in Medicine*, vol. 62, no. 1, pp. 1–10, 2014.
[276] J. A. Larson, M. H. Johnson, and S. B. Bhayani, "Application of surgical safety standards to robotic surgery: Five principles of ethics for nonmaleficence," *Journal of the American College of Surgeons*, vol. 218, no. 2, pp. 290–293, 2014.
[277] B. A. Knight, A. M. Potretzke, J. A. Larson, and S. B. Bhayani, "Comparing expert reported outcomes to national surgical quality improvement program risk calculator-predicted outcomes: Do reporting standards differ?" *Journal of Endourology*, vol. 29, no. 9, 2015.
[278] J. P. Halcox *et al.*, "Assessment of remote heart rhythm sampling using the AliveCor heart monitor to screen for atrial fibrillation," *Circulation*, vol. 136, pp. 1784–1794, 2017.
[279] M. P. Christiansen *et al.*, "Accuracy of a fourth-generation subcutaneous continuous glucose sensor," *Diabetes Technology & Therapeutics*, vol. 19, no. 8, 2017.
[280] J. Lawton *et al.*, "Patients' and caregivers' experiences of using continuous glucose monitoring to support diabetes self-management: qualitative study," *BMC Endocrine Disorders*, vol. 18, no. 12, 2018.
[281] E. Bruno *et al.*, "Wearable technology in epilepsy: The views of patients, caregivers, and healthcare professionals," *Epilepsy & Behavior*, vol. 85, pp. 141–149, 2018.
[282] G. Regalia, F. Onorati, M. Lai, C. Caborni, and R. W. Picard, "Multimodal wrist-worn devices for seizure detection and advancing research: Focus on the empatica wristbands," *Epilepsy Research*, vol. 153, pp. 79–82, 2019.
[283] E. R. Dorsey, A. M. Glidden, M. R. Holloway, G. L. Birbeck, and L. H. Schwamm, "Teleneurology and mobile technologies: the future of neurological care," *Nature Reviews Neurology*, vol. 14, pp. 285–297, 2018.
[284] M. Topalovic *et al.*, "Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests," *European Respiratory Journal*, vol. 53, no. 4, 2019.
[285] C. Delclaux, "No need for pulmonologists to interpret pulmonary function tests," *European Respiratory Journal*, vol. 54, no. 1, 2019.
[286] O. Niel, C. Boussard, and P. Bastard, "Artificial intelligence can predict GFR decline during the course of ADPKD," *American Journal of Kidney Diseases*, vol. 71, pp. 911–912, 2018.
[287] Y. J. Yang and C. S. Bang, "Application of artificial intelligence in gastroenterology," *World Journal of Gastroenterology*, vol. 25, no. 14, pp. 1666–1683, 2019.
[288] G. Campanella *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, pp. 1301–1309, 2019.
[289] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017.

[290] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, 2017.

[291] C. Chu *et al.*, "Applying machine learning to automated segmentation of head and neck tumour volumes and organs at risk on radiotherapy planning CT and MRI scans," *F1000 Research*, 2016.

[292] X. Liu *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The Lancet Digital Health*, vol. 1, no. 6, pp. e271–e297, 2019.

[293] C. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, 2019, DOI 10.1186/s12916-019-1426-2.

[294] P. Ding, M. Gurney, G. Perry, and R. Xu, "Association of COVID-19 with risk and progression of Alzheimer's Disease: Non-overlapping two-sample Mendelian randomization analysis of 2.6 million subjects," *Journal of Alzheimer's Disease*, vol. 96, no. 4, pp. 1711–1720, 2023.

[295] N. Naik *et al.*, "Legal and ethical considerations in artificial intelligence in healthcare: Who takes responsibility?" *Frontiers in Surgery*, vol. 9, 2022.

[296] K. H. Keskinbora, "Medical ethics considerations on artificial intelligence," *Journal of Clinical Neuroscience*, vol. 64, pp. 277–282, 2019.

[297] F. Howes, J. Doyle, N. Jackson, and E. Waters, "Evidence-based public health: The importance of finding "difficult to locate" public health and health promotion intervention studies for systematic reviews," *Journal of Public Health*, vol. 26, no. 1, pp. 101–104, 2004.

[298] M. Egger, G. Davey Smith, M. Schneider, and C. Minder, "Bias in meta-analysis detected by a simple, graphical test," *BMJ*, vol. 315, no. 7109, pp. 629–634, 1997.

[299] S. C. Hayes, J. Ciarrochi, S. G. Hofmann, F. Chin, and B. Sahdra, "Evolving an idionomic approach to processes of change: Towards a unified personalized science of human improvement," *Behaviour Research and Therapy*, vol. 156, 2022, DOI 10.1016/j.brat.2022.104155.

[300] S. C. Hayes *et al.*, "Report of the ACBS Task Force on the strategies and tactics of contextual behavioral science research," *Journal of Contextual Behavioral Science*, vol. 20, pp. 172–183, 2021.

[301] R. Johansson, T. Lofthouse, and P. Hammer, "Generalized identity matching in NARS," in *International Conference on Artificial General Intelligence*. Cham: Springer International Publishing, 2022, pp. 243–249.

[302] R. Johansson, P. Hammer, and T. Lofthouse, "Functional Equivalence with NARS," 2024. [Online]. Available: https://arxiv.org/abs/2405.03340

[303] R. Johansson, "Machine psychology: Integrating operant conditioning with the non-axiomatic reasoning system for advancing artificial general intelligence research," *Frontiers in Robotics and AI*, vol. 14, 2024, DOI 10.3389/frobt.2024.1440631.

[304] P. Wang, X. Li, and P. Hammer, "Self in NARS, an AGI System," *Frontiers in Robotics and AI*, vol. 5, 2018, Art. no. 20.

[305] K. Marx, *Das Kapital*. Verlag von Otto Meisner, Hamburg, 1867.

[306] J. Bengio *et al.*, "Managing extreme AI risks amid rapid progress," *Science*, vol. 384, no. 6698, pp. 842–845, 2024.

[307] W. Alharbi, "AI in the foreign language classroom: A pedagogical overview of automated writing assistance tools," *Education Research International*, 2023.

[308] M. Cascella, J. Montomoli, V. Bellini, and E. Bignami, "Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios," *Journal of Medical Systems*, vol. 47, no. 1, 2023, Art. no. 33.

[309] X. Hou, N. Omar, and J. Wang, "Interactive design psychology and artificial intelligence-based innovative exploration of anglo-american traumatic narrative literature," *Frontiers in Psychology*, vol. 12, 2021, Art. no. 755039.

[310] H. Kang and C. Lou, "AI agency vs. human agency: understanding human–AI interactions on TikTok and their implications for user engagement," *Journal of Computer-Mediated Communication: JCMC*, vol. 27, no. 5, 2022, Art. no. zmac014.

[311] T. Liu and X. Xiao, "A framework of AI-based approaches to improving ehealth literacy and combating infodemic," *Frontiers in Public Health*, vol. 9, 2021, Art. no. 755808.

[312] A. Rodríguez-Ruiz *et al.*, "Detection of breast cancer with mammography: Effect of an artificial intelligence support system," *Radiology*, vol. 290, no. 2, pp. 305–314, 2019.

[313] S. Kreps, R. Miles McCain, and M. Brundage, "All the news that's fit to fabricate: AI-generated text as a tool of media misinformation," *Journal of Experimental Political Science*, vol. 9, no. 1, pp. 104–117, 2022.

[314] B. C. Stahl, D. Schroeder, and R. Rodrigues, *Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges*. Springer Nature, 2022.

[315] C. Shao *et al.*, "The spread of low-credibility content by social bots," *Nature Communications*, vol. 9, no. 1, 2018, Art. no. 4787.

[316] D. Lazer *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[317] T. Caulfield, "Health misinformation and the power of narrative messaging in the public sphere," *Canadian Journal of Bioethics*, vol. 2, no. 2, pp. 52–60, 2019.

[318] K. M. Douglas, "COVID-19 conspiracy theories," *Group Processes & Intergroup Relations: GPIR*, vol. 24, no. 2, pp. 270–275, 2021.

[319] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[320] A. Bovet and H. A. Makse, "Influence of fake news in Twitter during the 2016 US presidential election," *Nature Communications*, vol. 10, no. 1, p. 7, 2019.

[321] A. Mitchell, M. Jurkowitz, J. Oliphant, B., and E. Shearer, "Most americans have heard of the conspiracy theory that the COVID-19 outbreak was planned, and about one-third of those aware of it say it might be true," Pew Research Center. [Online]. Available: https://www.pewresearch.org/journalism/2020/06/29/most-americans-have-heard-of-the-conspiracy-theory-that-the-covid-19-outbreak-was-planned-and-about-one-third-of-those-aware-of-it-say-it-might-be-true/

[322] Z. Lu, P. Li, W. Wang, and M. Yin, "The effects of AI-based credibility indicators on the detection and spread of misinformation under social influence," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSCW2, pp. 1–27, 2022.

[323] K. Shu *et al.*, "Fake news detection on social media: A data mining perspective," *SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, 2017.

[324] P. Akhtar *et al.*, "Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions," *Annals of Operations Research*, pp. 1–25, 2022.

[325] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, "Stop clickbait: Detecting and preventing clickbaits in online news media," 2016. [Online]. Available: https://arxiv.org/abs/1610.09786

[326] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11 765–11 788, 2021.

[327] N. L. Kolluri and D. Murthy, "CoVerifi: A COVID-19 news verification system," *Online Social Networks and Media*, vol. 22, p. 100123, 2021.

[328] S. McLennan *et al.*, "Embedded ethics: a proposal for integrating ethics into the development of medical AI," *BMC Medical Ethics*, vol. 23, no. 1, 2022, Art. no. 6.

[329] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, 2019.

[330] National Security Commission on Artificial Intelligence, "Final Report," Washington, D.C., Feb. 2021. [Online]. Available: https://www.nscai.gov/

[331] US Department of State, "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," Washington, D.C., Nov. 2023. [Online]. Available: https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-3/

[332] Raytheon Fact Sheet, "Phalanx Weapon System." [Online]. Available: https://www.rtx.com/raytheon/what-we-do/sea/phalanx-close-in-weapon-system

[333] K. Atherton, "Loitering munitions preview the autonomous future of warfare," Brookings Commentary. Washington, D.C., Aug. 2021. [Online]. Available: https://www.brookings.edu/articles/loitering-munitions-preview-the-autonomous-future-of-warfare/

[334] M. Sassoli, A. Bouvier, A. Quintin, and J. Grignon, "Lex Specialis," International Committee of the Red Cross, Geneva. [Online]. Available: https://casebook.icrc.org/a_to_z/glossary/lex-specialis

[335] V. Koutroulis, "Martens Clause," Oxford Bibliographies. Oxford, UK., Jul. 2013. [Online]. Available: https://www.oxfordbibliographies.com/display/document/obo-9780199796953/obo-9780199796953-0101.xml

[336] M. Sassoli, A. Bouvier, A. Quintin, and J. Grignon, "Fundamentals of IHL," International Committee of the Red Cross, Geneva. [Online]. Available: https://casebook.icrc.org/law/fundamentals-ihl

[337] ——, "Proportionality," International Committee of the Red Cross, Geneva. [Online]. Available: https://casebook.icrc.org/a_to_z/glossary/proportionality

[338] Human Rights Watch, "Killer Robots." [Online]. Available: https://www.hrw.org/topic/arms/killer-robots

[339] Yale Law School, "Annex to the Convention Regulations Respecting the Laws and Customs of War on Land, Section I, Article 1." The Hague, 18 October 1907. [Online]. Available: https://avalon.law.yale.edu/20th_century/hague04.asp

[340] M. Sassoli, A. Bouvier, A. Quintin, and J. Grignon, "Criminal Repression," International Committee of the Red Cross, Geneva. [Online]. Available: https://casebook.icrc.org/law/criminal-repression

[341] W. Knight, "The us wants china to start talking about ai weapons," Wired, New York, Nov. 2023. [Online]. Available: https://www.wired.com/story/us-china-killer-ai-weapons-apec-talks/

[342] V. Boulanin, "Implementing article 36 weapon reviews in the light of increasing autonomy in weapon systems," SIPRI Insights on Peace and Security, Nov. 2015. [Online]. Available: https://www.sipri.org/sites/default/files/files/insight/SIPRIInsight1501.pdf

[343] V. Boulanin and M. Verbruggen, "SIPRI compendium on article 36 reviews," SIPRI Background Paper, Stockholm, Dec. 2017. [Online]. Available: https://www.sipri.org/publications/2017/sipri-background-papers/sipri-compendium-article-36-reviews

[344] UN News, "UN and Red Cross call for restrictions on autonomous weapon systems to protect humanity," Oct. 2023. [Online]. Available: https://news.un.org/en/story/2023/10/1141922

[345] US Department of Defense press release, "DOD Adopts Ethical Principles for Artificial Intelligence," Washington, D.C., Feb. 2020. [Online]. Available: https://www.defense.gov/News/Releases/release/article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/

[346] E. Kania, "China's Strategic Ambiguity and Shifting Approach to Lethal Autonomous Weapons Systems," Lawfare Media, Apr. 17, 2018, Washington, D.C. [Online]. Available: https://www.lawfaremedia.org/article/chinas-strategic-ambiguity-and-shifting-approach-lethal-autonomous-weapons-systems

[347] K. Rogers and D. Sanger, "U.S. Manages Expectations of a Breakthrough Before Biden and Xi Meet," The New York Times, Nov. 14, 2023. [Online]. Available: https://www.nytimes.com/2023/11/14/us/politics/biden-xi-china-apec.html

[348] D. Kayser and A. Beck, "Convergence? European positions on lethal autonomous weapon systems," Pax for Peace report. Utrecht, The Netherlands, November 2019. [Online]. Available: https://paxvoorvrede.nl/wp-content/uploads/2020/11/convergence-pax-report-on-european-positions-on-lethal-autonomous-weapon-systems-update-2019.pdf

[349] United Nations, "First Committee Approves New Resolution on Lethal Autonomous Weapons, as Speaker Warns 'An Algorithm Must Not Be in Full Control of Decisions Involving Killing'," United Nations, New York, GA/DIS/3731, 1 Nov. 1 2023. [Online]. Available: https://press.un.org/en/2023/gadis3731.doc.htm

[350] D. Gilbert, "Here's How Violent Extremists Are Exploiting Generative AI Tools," Wired. New York, Nov. 3, 2023. [Online]. Available: https://www.wired.com/story/generative-ai-terrorism-content/

[351] M. Austero, A. F. Lubang, B. Nepram *et al.*, "Artificial Intelligence, Emerging Technology, and Lethal Autonomous Weapons Systems: Security, Moral, and Ethical Perspectives in Asia," Nonviolence International Southeast Asia, Manila. September 2020. [Online]. Available: https://www.stopkillerrobots.org/wp-content/uploads/2021/11/NISEA-AI-Emerging-Tech-and-LAWS-Perspectives-in-Asia.pdf

[352] G. Cooke, "Magic Bullets: The Future of Artificial Intelligence in Weapons Systems," US Army, Jun. 11, 2019. [Online]. Available: https://www.army.mil/article/223026/magic_bullets_the_future_of_artificial_intelligence_in_weapons_systems

[353] J. Metha, "The role of AI in website personalization," Abmatic AI, San Francisco, Nov. 18, 2023. [Online]. Available: https://abmatic.ai/blog/role-of-ai-in-website-personalization

[354] Golino, M.A., "Algorithms in Social Media Platforms. How social media algorithms influence the spread of culture and information in the digital society." Institute for internet & the Just Society. Apr. 24, 2021. [Online]. Available: https://www.internetjustsociety.org/algorithms-in-social-media-platforms

[355] Wu, K.J., "Radical ideas spread through social media. Are the algorithms to blame?" NOVA, March 28, 2019. [Online]. Available: https://www.pbs.org/wgbh/nova/article/radical-ideas-social-media-algorithms/

[356] R. Darbinyan, "How AI Transforms Social Media," Forbes, Mar. 16, 2023. [Online]. Available: https://www.forbes.com/councils/forbestechcouncil/2023/03/16/how-ai-transforms-social-media/

[357] J. Kaluža, "Habitual generation of filter bubbles: Why is algorithmic personalisation problematic for the democratic public sphere?" *Javnost - The Public*, vol. 29, pp. 1–17, 2021.

[358] L. Merten, "Block, hide or follow—personal news curation practices on social media," *Digital Journalism*, vol. 9, no. 8, pp. 1018–1039, 2021.

[359] J. A. Harriger, J. A. Eveans, J. K. Thompson, and T. L. Tylka, "The dangers of the rabbit hole: Reflections on social media as a portal into a distorted world of edited bodies and eating disorder risk and the role of algorithms," *Body Image*, vol. 41, pp. 292–297, 2022.

[360] S. Zeiger and J. Gyte, "Prevention of Radicalization on Social Media and the Internet," Handbook of Terrorism Prevention and Preparedness. International Centre for Counterterrorism. The Hague, 2021. [Online]. Available: https://www.icct.nl/sites/default/files/2023-01/Chapter-12-Handbook_0.pdf

[361] B. Mulin and N. Grant, "Google Tests A.I. Tool That Is Able to Write News Articles," The New York Times, 19 July 2023. [Online]. Available: https://www.nytimes.com/2023/07/19/business/google-artificial-intelligence-news-articles.html

[362] AIContentfy, "Exploring the Potential and Pitfalls of AI-Generated News Articles," October 29, 2024. [Online]. Available: https://aicontentfy.com/en/blog/potential-of-ai-generated-news-articles

[363] Verma, P., "The rise of AI fake news is creating a 'misinformation superspreader'," The Washington Post, Dec. 17, 2023. [Online]. Available: https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/

[364] Reuters, "OpenAI says AI tools can be effective in content moderation," August 15, 2023. [Online]. Available: https://www.reuters.com/technology/openai-says-ai-tools-can-be-effective-content-moderation-2023-08-15/

[365] Darbinyan, R., "The Growing Role Of AI In Content Moderation," Forbes, Jun. 24, 2022. [Online]. Available: https://www.forbes.com/councils/forbestechcouncil/2022/06/14/the-growing-role-of-ai-in-content-moderation/

[366] R. Spence, A. Bifulco, P. Bradbury, E. Martellozzo, and J. DeMarco, "The psychological impacts of content moderation on content moderators: A qualitative study," *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, vol. 17, no. 4, 2023.

[367] Newton, C., "The Trauma Floor," The Verge, Feb. 25, 2019. [Online]. Available: https://www.theverge.com/2019/2/25/18229714/cognizantfacebook-content-moderator-interviews-trauma-working-conditionsarizona

[368] Paul, C. and Reininger, H., "Platforms Should Use Algorithms to Help Users Help Themselves," Carnegie Endowment, Jul. 20, 2021. [Online]. Available: https://carnegieendowment.org/posts/2021/07/platforms-should-use-algorithms-to-help-users-help-themselves?lang=en

[369] Barret, P., "It's Past Time to Take Social Media Content Moderation In-House," Just Security, Jan. 18, 2023. [Online]. Available: https://www.justsecurity.org/84812/its-past-time-to-take-social-mediacontent-moderation-in-house/

[370] Marvin, R., "Facebook's AI Is Good at Detecting Spam but Struggles With Hate Speech and Harassment," PC Magazine, Jun. 3, 2019. [Online]. Available: https://www.pcmag.com/news/facebooks-ai-is-good-at-detecting-spam-but-struggles-with-hate-speech-and

[371] Walsh, D., "Amy Zegart: Integrating AI in the Realm of National Security," Stanford, California, Nov. 7, 2023. [Online]. Available: https://hai.stanford.edu/news/amy-zegart-integrating-ai-realm-national-security

[372] Rigano, C., "Using Artificial Intelligence to Address Criminal Justice Needs," NIJ Journal 280, January 2019. [Online]. Available: https://www.ojp.gov/pdffiles1/nij/252038.pdf

[373] M. Parsapoor, J. W. Koudys, and A. C. Ruocco, "Suicide risk detection using artificial intelligence: the promise of creating a benchmark dataset for research on the detection of suicide risk," *Frontiers in Psychiatry*, vol. 14, 2023, DOI 10.3389/fpsyt.2023.1186569.

[374] E. Fosch-Villaronga and G. Malgieri, "Queering the ethics of AI," in *Handbook on the Ethics of Artificial Intelligence*, D. J. Gunkel, Ed. E. Elgar Publishing, 2024.

[375] S. Sulmicelli, "Algorithmic content moderation and the LGBTQ+ community's freedom of expression on social media: Insights from the EU

digital services act," *BioLaw Journal - Rivista Di BioDiritto*, vol. 2, pp. 471–489, 2023.

[376] Network Contagion Research Institute, "A Tik-Tok-ing Timebomb: How TikTok's Global Platform Anomalies Align with the Chinese Communist Party's Geostrategic Objectives," Dec. 2023. [Online]. Available: https://networkcontagion.us/wp-content/uploads/A-Tik-Tok-ing-Timebomb_12.21.23.pdf

[377] Miller, K., "Privacy in an AI Era: How Do We Protect Our Personal Information?" Human Centered Artificial Intelligence. Stanford University, Mar. 18, 2024. [Online]. Available: https://hai.stanford.edu/news/privacy-ai-era-how-do-we-protect-our-personal-information

[378] United Nations, "Data Protection and Privacy Legislation Worldwide," UN Trade and Development. [Online]. Available: https://unctad.org/page/data-protection-and-privacy-legislation-worldwide

[379] X. Chen, D. Zou, H. Xie, G. Cheng, and C. Liu, "Two decades of artificial intelligence in education," *Educational Technology & Society*, vol. 25, no. 1, pp. 28–47, 2022.

[380] D. McArthur, M. Lewis, and M. Bishary, "The roles of artificial intelligence in education: current progress and future prospects," *Journal of Educational Technology*, vol. 1, no. 4, pp. 42–80, 2005.

[381] V. Devedžić, "Web intelligence and artificial intelligence in education," *Journal of Educational Technology & Society*, vol. 7, no. 4, pp. 29–39, 2004.

[382] E. Kasneci *et al.*, "ChatGPT for good? on opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, 2023, DOI 10.1016/j.lindif.2023.102274.

[383] X. Zhai *et al.*, "A review of artificial intelligence (AI) in education from 2010 to 2020," *Complexity*, pp. 1–1, 2021.

[384] I. Straw and C. Callison-Burch, "Artificial Intelligence in mental health and the biases of language-based models," *PloS One*, vol. 15, no. 12, 2020, Art. no. e0240376.

[385] S. C. Hayes, J. Ciarrochi, S. G. Hofmann, F. Chin, and B. Sahdra, "Evolving an idionomic approach to processes of change: Towards a unified personalized science of human improvement," *Behaviour Research and Therapy*, vol. 156, 2022, Art. no. 104155.

[386] S. Bhattacharya, C. Goicoechea, S. Heshmati, J. K. Carpenter, and S. G. Hofmann, "Efficacy of cognitive behavioral therapy for anxiety-related disorders: A meta-analysis of recent literature," *Current Psychiatry Reports*, vol. 25, no. 1, pp. 19–30, 2023.

[387] A. C. Timmons *et al.*, "A call to action on assessing and mitigating bias in artificial intelligence applications for mental health," *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, vol. 18, no. 5, pp. 1062–1096, 2023.

[388] E. Feczko, O. Miranda-Dominguez, M. Marr, A. M. Graham, J. T. Nigg, and D. A. Fair, "The heterogeneity problem: Approaches to identify psychiatric subtypes," *Trends in Cognitive Sciences*, vol. 23, no. 7, pp. 584–601, 2019.

[389] J. Ciarrochi, B. Sahdra, S. C. Hayes, S. G. Hofmann, B. Sanford, C. Stanton, K. Yap, M. I. Fraser, K. Gates, and A. T. Gloster, "A personalized approach to identifying important determinants of well-being," *Cognitive Therapy and Research*, vol. 48, pp. 1–22, 2024.

[390] M. C. Fadus, "Unconscious bias and the diagnosis of disruptive behavior disorders and ADHD in African American and Hispanic youth," *Academic Psychiatry: The Journal of the American Association of Directors of Psychiatric Residency Training and the Association for Academic Psychiatry*, vol. 44, no. 1, pp. 95–102, 2020.

[391] H. N. Garb, "Race bias and gender bias in the diagnosis of psychological disorders," *Clinical Psychology Review*, vol. 90, 2021, Art. no. 102087.

[392] T. A. Widiger and R. L. Spitzer, "Sex bias in the diagnosis of personality disorders: Conceptual and methodological issues," *Clinical Psychology Review*, vol. 11, no. 1, pp. 1–22, 1991.

[393] S. Riches *et al.*, "Therapeutic engagement in robot-assisted psychological interventions: A systematic review," *Clinical Psychology & Psychotherapy*, vol. 29, no. 3, pp. 857–873, 2022.

[394] L. B. Dixon, Y. Holoshitz, and I. Nossel, "Treatment engagement of individuals experiencing mental illness: review and update," *World Psychiatry: Official Journal of the World Psychiatric Association*, vol. 15, no. 1, pp. 13–20, 2016.

[395] S. Barello *et al.*, "eHealth for patient engagement: A systematic review," *Frontiers in Psychology*, vol. 6, pp. 1–15, 2013.

[396] R. H. Christie, A. Abbas, and V. Koesmahargyo, "Technology for measuring and monitoring treatment compliance remotely," *Journal of Parkinson's Disease*, vol. 11, no. s1, pp. S77–S81, 2021.

[397] L. Chaby *et al.*, "Embodied virtual patients as a simulation-based framework for training clinician-patient communication skills: An overview of their use in psychiatric and geriatric care," *Frontiers in Virtual Reality*, vol. 3, 2022, Art. no. 827312.

[398] R. Johansson, G. Skantze, and A. Jönsson, "A psychotherapy training environment with virtual patients implemented using the Furhat robot platform," in *Intelligent Virtual Agents: 17th International Conference, Proceedings*, vol. 17. Springer International Publishing, 2017, pp. 184–187.

[399] A. Nye, J. Delgadillo, and M. Barkham, "Efficacy of personalized psychological interventions: A systematic review and meta-analysis," *Journal of Consulting and Clinical Psychology*, vol. 91, no. 7, pp. 389–397, 2023.

[400] R. Daryabeygi-Khotbehsara *et al.*, "Smartphone-based interventions to reduce sedentary behavior and promote physical activity using integrated dynamic models: Systematic review," *Journal of Medical Internet Research*, vol. 23, no. 9, 2021, Art. no. e26315.

[401] S. J. Beard and R. Bronson, "The story so far: How humanity avoided existential catastrophe," in *Cambridge Conference on Catastrophic Risk 2020*, 2022.

[402] M. Graves, "Why we should be concerned about artificial superintelligence," Skeptic, 2017. [Online]. Available: https://www.skeptic.com/reading_room/why-we-should-be-concerned-about-artificial-superintelligence/

[403] S. Russell, *Human Compatible – Artificial Intelligence and the Problem of Control*. Viking (US), 2019.

[404] D. Hendrycks, M. Mazeika, and T. Woodside, "An overview of catastrophic AI risks," 2023. [Online]. Available: https://arxiv.org/abs/2306.12001

[405] H. Belfield and C. Ruhl, "Why policy makers should beware claims of new ärms races," Bulletin of the Atomic Scientists, 2022. [Online]. Available: https://thebulletin.org/2022/07/why-policy-makers-should-beware-claims-of-new-arms-races/

[406] M. Maas, K. Matteucci, and D. Cooke, "Military artificial intelligence as contributor to global catastrophic risk," *The Era of Global Risk*, 2023.

[407] E. Karger, P. Atanasov, and P. Tetlock, "Improving judgments of existential risk: Better forecasts, questions, explanations, policies," 2022.

[408] R. Binzel, "The Torino impact hazard scale," *Planetary and Space Science*, vol. 48, no. 4, pp. 297–303, 2000.

[409] Quantified Strategies, "Top AI Sentiment Analysis for Trading Strategies," Sep. 2024. [Online]. Available: https://www.quantifiedstrategies.com/ai-sentiment-analysis-for-trading/

[410] P. Weights, "AI-Powered Portfolio Optimization: the future of Asset Management?" Sep. 2023. [Online]. Available: https://swissfinte.ch/ai-powered-portfolio-optimization-future-of-asset-management/

[411] Q. Cheng, L. Yang, J. Zheng, M. Tian, and D. Xin, "Optimizing portfolio management and risk assessment in digital assets using deep learning for predictive analysis," 2024. [Online]. Available: https://arxiv.org/abs/2402.15994

[412] M. Ford, *Rise of the Robots: Technology and the Threat of a Robotics Future*. Basic Books (US), 2015.

[413] ——, *Rule of the Robots: How Artificial Intelligence Will Transform Everything*. Hachette (UK), 2021.

[414] G. Kallis, F. Demaria, and G. D'Alisa, *Degrowth: Vocabulary for a New Era*. Routledge, New York, 2015.

[415] A. Barrau, *Le plus grand défi de l'histoire de l'humanité*. Michel Lafon, 2020.

[416] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583–589, 2021.

[417] J. Wang *et al.*, "Clinical information extraction applications: A literature review," *Journal of Biomedical Informatics*, vol. 77, pp. 34–49, 2018.

[418] P. Jensen, L. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, pp. 395–405, 2012.

TOMMASO DORIGO was born in Venice, Italy in 1966. He received his Laurea in physics in 1995 and a Ph.D. in physics in 1999, both from the University of Padova.

From 1999 to 2000, he was a postdoctoral fellow at Harvard University and later a research grantee with the University of Padova and an INFN researcher. Since 2019 he is a First Researcher at the Padova section of INFN (Italian National Institute for Nuclear Physics), and since 2024 he is also a WASP Guest Researcher at Luleå University of Technology (Sweden).

Dr. Dorigo is currently serving as President of the USERN organization (https://usern.org). He is also the founder of the MODE Collaboration (https://mode-collaboration.github.io). He was a member of the CDF experiment at Fermilab from 1996 to 2012 and is a member of the CMS experiment at CERN since 2001. He is an author of over 1700 scientific publications in peer-reviewed journals and has an H-index of 257, according to Google Scholar.

GARY D. BROWN is an associate professor of practice and faculty lead for the cyber policy concentration at the Bush School of Government and Public Service, Texas A&M University.

Previously, he taught cyber policy and law at the US National Defense University and Marine Corps University. He served for 24 years as an officer in the US Air Force, retiring as a colonel. After his military career, he spent three years with the International Committee of the Red Cross Washington Delegation.

Colonel Brown was the first senior legal counsel for US Cyber Command, has authored a number of cyber policy and law articles, and was the on-camera legal expert for the documentary film Zero Days (2016). He has a law degree from the University of Nebraska and an LL.M. in international law from Cambridge University.

CARLO CASONATO is a Full Professor of comparative constitutional law at the Faculty of Law of the University of Trento.

After receiving a Fellowship from the Council of Europe (1993), he completed a PhD in Fundamental Rights and Freedoms in Comparative Law (1996). Visiting professor at the Chicago-Kent College of Law (IIT, 2003: course on Law and Bioethics), he has held teaching appointments or research periods at the Universities of Oxford (2016, 2013, and 2012), Harvard (2009), Berkeley (2006), Toronto (2000), Lancaster (1999), Montréal (1995) and San Sebastian (1993). He is the Principal Investigator for numerous national and European research projects, and he is the author or editor of over 160 publications, including more than 20 books.

Prof. Casonato holds the Jean Monnet Chair in AI EU Law (T4F). He is the founder and chief editor of the BioLaw Journal, Director of the BioLaw Laboratory, and serves as the delegate of the rector and vice-president of the Ethics Committee for Research at the University of Trento. He is also a member of the Commission for Ethics and Integrity in Research at the CNR (Italian National Research Council).

ARTEMI CERDÀ is a Full Professor in Physical Geography at the University of València. Head of the Soil Erosion and Degradation Research Group.

He developed his Ph.D. in the soil hydrology of Mediterranean Soils (1989-1993). During his post-doctorate studies in The Netherlands, Israel, Bolivia, and the UK from 1994 to 1996, he researched (i) soil hydrology in arid land, (ii) the impact of climate on geomorphological processes, and (iii) fire and agriculture impact on soil sustainability. His position as a researcher in the Desertification Research Centre shed light on the interaction of seeds, plants, and soil erosion in road embankments.

Since 2002, Prof. Cerdá has worked as an associate professor, got his accreditation as a Full Professor in 2009, and has worked since 2010 at the University of Valencia as a full professor. He developed the Soil Erosion and Degradation Research Group (SEDER) and the Soil Erosion Research Stations of Montesa and El Teularet. These studies focused mainly on soil erosion processes in agriculture and forest land.

JOSEPH CIARROCHI , PhD, is a professor at the Institute for Positive Psychology and Education at Australian Catholic University.

He has published more than 179 scientific journal articles and many books, including the widely acclaimed Emotional Intelligence in Everyday Life and The Weight Escape. His latest book is What makes you stronger: How to thrive in the face of uncertainty using Acceptance and Commitment Therapy.

Prof. Ciarrochi has served as the President of the Association for Contextual Behavior Science and was the first editor of the Journal of Contextual Behavioral Science. Prof. Ciarrochi has been honored with more than four million dollars in research funding. His work has been discussed on TV and radio, and in magazines and newspaper articles. He is ranked in the top 1% of scientists in the world across all disciplines.

MAURO DA LIO is a Full Professor of mechanical systems at the University of Trento, Italy.

His earlier research activity was modeling, simulation, and optimal control of mechanical multibody systems, particularly vehicle and spacecraft dynamics. More recently, his focus shifted to modeling human sensory-motor control with applications in health, robotics, and, mostly, intelligent vehicles. He was involved in several EU framework program 6 and 7 projects (PReVENT, SAFERIDER, interactIVe, VERITAS, AdaptIVe, No-Tremor, and SUNRISE).

Prof. Da Lio was the coordinator of the EU Horizon 2020 Dreams4Cars Research and Innovation Action: a collaborative project in the robotics domain that aimed at increasing the cognition abilities of artificial driving agents using offline simulation mechanisms broadly inspired by the human dream state (learning of forward models and offline synthesis of inverse ones).

NICOLE D'SOUZA is a researcher at the Systems Neural Engineering Laboratory at the University of California, Riverside. She is pursuing a Bachelor of Science in Neuroscience while also minoring in Data Science. Her academic journey is marked by a commitment to bridging the gap between neuroscience and data analytics.

She is the recipient of the prestigious Chancellor's Research Fellowship, funding her research between the Department of Bioengineering, UCR and Stanford University. Currently, she is heading research at the Palo Alto Veterans Affair Medical Center and Stanford in applying machine learning algorithms to enhance diagnostic tools and treatment strategies for neurological disorders such as Alzheimer's Disease.

Ms. D'Souza serves as an editor in UCR's Undergraduate Research Journal, where she contributes to the dissemination of innovative research findings, and is a previous intern at the Society of Brain Mapping and Therapeutics, where she organized the Alzheimer's Disease and Dementia conference track.

NICOLAS R. GAUGER is Full Professor and Chairholder for Scientific Computing and Director of the Computing Center (RHRZ) at University of Kaiserslautern-Landau (RPTU) as well as Principal Investigator of the SIVERT research training group (https://sivert.info) dealing with the algorithmic part of the proton Computed Tomography (pCT) project of the Bergen pCT Collaboration (https://indico.cern.ch/category/13882/).

His research interests are numerical optimization, high-performance computing, machine learning, and pCT, amongst other fields of application.

Prof. Gauger is also a member of the MODE Collaboration (https://mode-collaboration.github.io), a collaboration of physicists and computer scientists for the optimized co-design of hardware and software for future experiments in fundamental science.

STEVEN C. HAYES received his Ph.D. in clinical psychology with a minor in experimental psychology from West Virginia University, Morgantown, WV, USA, in 1977, doing his internship at Brown University, Providence, RI in 1975-76.

He was an Assistant and Associate Professor of Psychology from 1977 to 1986 at the University of North Carolina at Greensboro, and from 1986 to 2023, a Full Professor of Psychology at the University of Nevada, Reno. He is now Foundation Professor Emeritus at UNR and President of the Institute for Better Health in Santa Rosa, CA, a 47-year-old charitable organization dedicated to excellence in mental and behavioral health care. As the author of 47 books and over 700 scientific articles, his career has focused on symbolic learning and how it relates to human suffering.

Dr. Hayes is a Fellow of the American Association for the Advancement of Science and has received lifetime achievement awards from the Association for Behavioral and Cognitive Therapy, and the Association for Psychological Science.

STEFAN G. HOFMANN was born in Germany. He is the Alexander von Humboldt Professor at the Philipps University of Marburg (Germany), where he holds the LOEWE Spitzenprofessur for Translational Clinical Psychology (Germany). Hofmann studied psychology at the University of Marburg, Germany, where he received his B.A., M.S., and Ph.D.

He went to the USA in 1991 and worked at Boston University from 1996 to 2023. Prior to that, he was at SUNY Albany and, before that, at Stanford University. Hofmann is a leading expert in anxiety disorders, with a focus on psychotherapy. He has been the editor of Cognitive Therapy and Research and has been named a Highly Cited Researcher every year since 2015.

Prof. Hofmann is a fellow of many professional organizations, including AAAS, and he has received many awards, including the Aaron T. Beck Award for Significant and Enduring Contributions to the Field of Cognitive Therapy, the Lifetime Achievement Award by ABCT, and the Humboldt Research Award. He was an advisor to the DSM-5 and the DSM-5-TR. He has published more than 500 peer-reviewed scientific articles and 20 books. Some of his studies led to insights into the mechanism of treatment change, translating discoveries from neuroscience into clinical applications, emotion regulation, and cultural expressions of psychopathology.

ROBERT JOHANSSON is an Associate Professor in Clinical Psychology in the Department of Psychology at Stockholm University.

He holds a PhD (2013) in Psychology and a PhD (2024) in Computer Science, both from Linköping University. His clinical research has focused on the development and evaluation of digital health interventions. Since 2017, his main focus of research has been in the area of Artificial General Intelligence (AGI), where he has focused on studying the development of human-level cognitive capabilities with a particular AGI system, the Non-Axiomatic Reasoning System (NARS). The research is guided by Relational Frame Theory, a behavioral psychology approach to understanding human language and cognition. He is also a licensed psychologist with a broad range of interests, with a particular interest in emotion-focused psychotherapy models.

MARCUS LIWICKI is chaired professor in Machine Learning and vice-rector for AI at Luleå University of Technology. He received his M.S. degree in Computer Science from the Free University of Berlin, Germany, in 2004, his PhD degree from the University of Bern, Switzerland, in 2007, and his habilitation degree at the Technical University of Kaiserslautern, Germany, in 2011.

His research interests include machine learning, pattern recognition, artificial intelligence, human-computer interaction, digital humanities, knowledge management, ubiquitous intuitive input devices, document analysis, and graph matching.

From October 2009 to March 2010, Prof. Liwicki visited Kyushu University (Fukuoka, Japan) as a research fellow (visiting professor), supported by the Japanese Society for the Promotion of Science. In 2015, at the young age of 32, he received the ICDAR Young Investigator Award, a bi-annual award acknowledging outstanding achievements in pattern recognition for researchers up to the age of 40.

FABIEN LOTTE is a research director (DR2) at the Inria Center at the University of Bordeaux and the LaBRI, Talence, France.

He holds an MSc. (INSA Rennes, 2005), a PhD (INSA Rennes, 2008), and a Habilitation to Supervise Research (Univ. Bordeaux, 2016), all in Computer Science. Fabien Lotte is a specialist in Brain-Computer Interfaces (BCI) and ElectroEncephaloGraphy signal processing. He is a member of the editorial boards of several leading journals on BCI and co-edited two books on the subject in 2016 and 2018.

Dr. Lotte notably coordinated or is coordinating the ANR REBEL project (2016-2019), the ANR Proteus project (2023-2027), the ERC Starting Grant BrainConquest project (2017-2022) or the ERC Proof-of-Concept project SPEARS (2024-2025). He has published more than 200 papers in this field, which have received more than 16 000 citations in total. He also gave more than 110 invited talks in 20 different countries. He is the laureate of the 2022 international USERN (Universal Science and Education Research Network) prize in formal science, the 2023 Lovelace-Babbage prize from the French Academy of Science in collaboration with the French Computer Science Society (SIF), and the Nature Award for Mentoring in Science 2023 (mid-career category).

JUAN J. NIETO is a Full Professor at the Department of Statistics, Mathematical Analysis and Optimization and Academician. He was a Fulbright fellow at the University of Texas (USA), and his research areas include Mathematical Analysis, Differential Equations, Nonlinear Analysis, Biomedical Applications, and Digital Twins.

His most influential contributions to date are in the area of differential equations, and his research interests are in fractional calculus, equations under uncertainty, and epidemiological models. He is one of the most cited mathematicians in the world according to different databases and has been listed in the Highly Cited Researchers uninterruptedly from 2014 to 2021. He was also in the World's Top 2% Researcher in 2022 by Stanford University. In 2010, he was among the scientists who had the most hot papers.

Dr. Nieto is presently Editor-in-Chief of the journal Fixed Point Theory and Algorithms for Sciences and of Differential Equations and Dynamical Systems.

GIULIA OLIVATO is a third-year PhD student at the University of Trento. In 2020, she graduated cum laude in Comparative, European and Transnational Law at the University of Trento.

She was a visiting PhD student at Technische Universität München (Germany) from October 2023 to April 2024. She completed a 6-month internship (November 2022 - June 2023) at a software company, where she experienced first-hand the benefits and drawbacks of implementing some regulatory provisions of the AI Act.

Ms. Olivato received formal legal training in both Rome and Trento and has been admitted to the bar. She has presented at various conferences and webinars and has published one article and two conference proceedings. Additionally, one more conference proceeding and one book chapter are forthcoming.

PETER PARNES was born in Malmö, Sweden, in 1971. He earned his PhD in computer science from Luleå University of Technology (LTU), Luleå, Sweden, in 1999. He holds there a docent title in media technology.

He has been a Full Professor in Pervasive and Mobile Computing at Luleå University of Technology since 2010, leading the ArcTech Learning Lab since 2020. His work experience includes various academic and research positions, including his role as a researcher and lecturer at LTU before his current position, and several non-academic positions, such as founder and chief scientist of Marratech 1998-2007 and site engineering manager for Google Sweden 2007-2009.

Prof. Parnes has received the Curt Boström award for best PhD thesis (1999) and IT-person of the year (2015) for his ongoing work in attracting more girls into the STEM area and technical studies via his work with Luleå Makerspace. He has served as an expert in various subjects related to digitalization and learning with several Swedish authorities, such as the Swedish National Agency for Education, the Swedish Ethical Review Authority, the Swedish Research Council, and Sweden's innovation agency, VINNOVA.

GEORGE PERRY is a neuroscientist and Professor of Neuroscience, Developmental and Regenerative Biology at the University of Texas at San Antonio. Perry is recognized in the field of Alzheimer's disease research, particularly for his work on oxidative stress.

He received his Bachelor of Arts degree in zoology from the University of California, Santa Barbara. After graduation, he studied at Scripps Institution of Oceanography, Hopkins Marine Station of Stanford University, and the Marine Biological Laboratory at Woods Hole; he obtained his Ph.D. from the University of California at San Diego in Marine Biology under David Epel in 1979. He then received a postdoctoral fellowship in the Department of Cell Biology in the laboratories of William R. Brinkley, Joseph Bryan, and Anthony R. Means at Baylor College of Medicine, where he laid the foundation for his observations of cytoskeletal abnormalities.

Prof. Perry joined in 1982 the faculty of Case Western Reserve University, where he holds an adjunct appointment. He is dean of the College of Sciences and professor of biology at the University of Texas at San Antonio. He is distinguished as one of the top Alzheimer's disease researchers with over 1000 publications, one of the top 100 most-cited scientists in Neuroscience & Behavior, and one of the top 25 scientists in free radical research.

ALICE PLEBE received her BSc and MSc in Computer Science from the University of Catania, Italy, in 2014 and 2016, respectively. In 2021, she earned her Ph.D. in Information and Communication Technology from the University of Trento, Italy.

Following her doctorate, she was a postdoctoral fellow at the same institution from 2021 to 2024. As part of the EU Horizon 2020 Dreams4Cars Research and Innovation Action, her research focused on exploring how the cognitive capabilities of the human brain can inform the design of artificial driving agents with human-like performance.

Dr. Plebe since 2024 has been a research fellow at University College London, working on large language models for trustworthy collaboration in multi-agent systems.

IDUPULAPATI M. RAO was born in Mandadam Village, Amaravathi, Andhra Pradesh, India in 1951.

He received a B.Sc. degree in Chemistry, Botany, and Zoology in 1971 from Andhra University; an M.Sc. degree in Botany in 1973 from Bhopal University; and a Ph.D. degree in Botany (Plant Physiology) in 1978 from Sri Venkateswara University in India.

From 1979 to 1981, Dr. Rao was a Plant Physiologist with ICRISAT, India. From 1981 to 1989, he was a Research Associate/Assistant Specialist at the University of Illinois and the University of California-Berkeley, USA. From 1989 to 2016, he worked as a Plant Nutritionist/Physiologist at the International Center for Tropical Agriculture (CIAT), Cali, Colombia. Since 2017, he has been an International Consultant and Emeritus Scientist of CIAT. He is the author of more than 250 journal articles and more than 70 book chapters. His research interests include abiotic stress tolerance of crops, crop-livestock systems, and mitigation of climate change. He is on the Editorial Board of the journal Farming System. Dr. Rao was a recipient of the Outstanding Principal Staff Award in 2000 from CIAT and the Outstanding Research Publication Awards from CIAT in 1999, 2003, 2009, and 2011.

NIMA REZAEI gained his medical degree (MD) from Tehran University of Medical Sciences and subsequently obtained an MSc in Molecular and Genetic Medicine and a PhD in Clinical Immunology and Human Genetics from the University of Sheffield, UK. He also spent a short-term fellowship in Pediatric Clinical Immunology and Bone Marrow Transplantation at Newcastle General Hospital.

He is now the Full Professor of Immunology and Vice Dean of Research and Technologies at the School of Medicine, Tehran University of Medical Sciences, and the cofounder and Head of the Research Center for Immunodeficiencies. He is also the Founder of the Universal Scientific Education and Research Network (USERN).

Prof. Rezaei has already been the Director of more than two hundred research projects and has designed and participated in several international collaborative projects. Prof. Rezaei is the editor, editorial assistant, or editorial board member of more than fifty international journals. He has edited more than one hundred international books, has presented more than a thousand lectures/posters at congresses/meetings, and has published more than 1,500 scientific papers in international journals.

FREDRIK SANDIN was born in Sweden in 1977. He is a full professor in machine learning at the Luleå University of Technology (LTU) since 2021.

He was an NFSR postdoctoral fellow at the University of Liège 2008-2009 and received his Ph.D in Physics from LTU and the Swedish Graduate School of Space Technology in 2007, both with a focus on computational physics. He completed his M.Sc. in 2001 with a diploma work in the ATLAS Collaboration at CERN.

His work focuses on brain-inspired machine learning and neuromorphic technologies for resource-efficient AI. He serves as PI for several projects in that area and coordinates the LTU research and education on sustainable AI. He was awarded the Gunnar Öquist Fellowship from the Kempe Foundations in 2014 and received a "New-Talents" award for original work in theoretical physics at the 2004 International School of Subnuclear Physics in Erice.

ANDREY USTYUZHANIN is the Director of AI/ML research at Acronis (Singapore), a visiting research professor at the National University of Singapore, and a PI at the Institute of Functional Intelligent Materials. His research's primary priority is designing new Machine Learning methods and using them to solve tough scientific challenges, thus improving the fundamental understanding of our world.

His current research interest is the development of foundation generative models for the discovery of new dynamic materials. Before joining NUS, he worked with LHCb – one of the LHC experiments at CERN.

Dr. Ustyuzhanin's key projects include the efficiency improvement of online triggers at LHCb, speeding up BDT-based online processing, and developing the algorithm for tracking in scintillators optical fiber detectors, among others.

GIORGIO VALLORTIGARA is a Professor of Neuroscience at the Centre for Mind/Brain Sciences of the University of Trento, Italy.

He has published more than 300 refereed papers that have received more than 29,000 citations (h-index=95 Google Scholar).

Prof. Vallortigara has also contributed to several book chapters and is the author of "Born Knowing" (MIT Press, 2021). He has been the recipient of the Geoffrey de St. Hilaire Prize for Ethology and a doctorate honoris causa from the University of Ruhr in Germany.

PIETRO VISCHIA was born in Padova, Italy in 1983. After his Bachelor (2006) and Master (2011) in physics from Università degli Studi di Padova, he obtained his Ph.D. in Physics in 2016 from Instituto Superior Técnico, Lisboa.

From 2016 to 2018, he was a postdoctoral fellow at Universidad de Oviedo and from 2018 to 2022 at Université catholique de Louvain, holding several prestigious doctoral and postdoctoral research grants. Since 2023, he is in his current position as a "Ramón y Cajal" Senior Researcher at Universidad de Oviedo and ICTEA and Adjunct Professor at IIT-Madras.

He is the coordinator and one of the cofounders of the MODE Collaboration (mode-collaboration.github.io), a collaboration of physicists and computer scientists for the optimized co-design of hardware and software for future experiments in fundamental science. He is also coordinator of the Machine Learning group of the CMS Collaboration, and he was the co-coordinator of the CERN Interexperimental Machine Learning Working Group from 2020 to 2024. Prof. Vischia has been a member of the CMS Collaboration at CERN since 2009. He is an author of over 1000 scientific publications in peer-reviewed journals and has an H-index of 116, according to Scopus. More on his research can be found at vischia.github.io/.

NILOUFAR YAZDANPANAH is a researcher in immunology, neuroimmunology, and genetics. She received her MD from Tehran University of Medical Sciences.

In 2022, she received the title of "Best Undergraduate Student of the Year" at the Tehran University of Medical Sciences and also received the same award at Children's Medical Center Hospital; in 2023, she was awarded the First-Rank Razi Medical Award, the most prestigious award in medical sciences in Iran.

Currently, Dr. Yazdanpanah is a post-doc researcher at the Research Center for Immunodeficiencies, Children's Medical Center of Tehran. Dr. Yazdanpanah has been the Executive Director of the Universal Scientific Education and Research Network (USERN) since 2022.