Artificial Intelligence: Powering the Fourth Industrial Revolution[1]

Angelo Cangelosi (University of Manchester) & David Vernon (Carnegie Mellon University Africa)

Abstract Artificial Intelligence (AI) is the branch of computer science and engineering that allows us to harness the power of computing and technology to mimic and extend human intelligence. Together with ubiquitous communications and near-universal access to information, artificial intelligence is driving the Fourth Industrial Revolution, ushering in an era of unprecedented and rapid change in how humans live, work, and relate to one another through the fusion of physical, digital, and biological technologies. In this article, we trace the origin and evolution of the different strands of AI and consider the implications of its pervasive presence in society, addressing some of its many applications — in medicine, robotics, the world-wide web & social media, and sport — and their impact on society across the globe, in developed and developing countries, and the ethical issues it raises for humankind.

Keywords: Artificial intelligence, cognitive science, fourth industrial revolution, deep learning, neural networks, symbolic computing, social impact, ethics.

1. What is AI, where did it come from, and where is it taking us?

In 1960, J. C. R. Licklider predicted a symbiotic partnership between humans and computers that will perform intellectual operations much more effectively than humans alone can perform them (Licklider, 1960). Today, that symbiotic partnership is now being realized through AI, a technology that both amplifies and extends human cognitive abilities. While there is some concern today about the ultimate destiny of that partnership and whether or not AI will prevail over humans if we reach the technological singularity when the autonomous capabilities of AI exceed those of humans (Shanahan, 2015), people are also concerned about how to harness AI for economic advantage and social development. From this latter perspective, AI forms the foundation of the fourth industrial revolution, a revolution that is characterized by a fusion of physical, digital, and biological technologies, powered by AI and enabled by ubiquitous communication and near-universal access to information. It is irreversibly altering how humans live, work, and relate to one another (Schwab, 2021). At the same time, it is important to consider how to harness AI within an ethical framework that achieves economic benefits and social development for all.

The world of the AI-powered fourth industrial revolution may well be the destination, but how did AI get started? For many people, the discipline of AI has its origins in a conference held at Dartmouth College, New Hampshire, in July and August 1956. It was attended by luminaries such as John McCarthy (who coined the term artificial intelligence; see Fig. 1), Marvin Minsky, Allen Newell, Herbert Simon, and Claude Shannon, all of

[1] A short version of this article has been accepted for publication in *EPS Grand Challenges*: *Physics for Society at the Horizon 2050* coordinated by the European Physical Society.

whom had a very significant influence on the development of AI over the next half-century. The essential position of AI at this time was that intelligence — both biological and artificial — is achieved by computations performed on internal symbolic knowledge representations, an approach referred to as *computationalism*, grounded in cognitivist psychology, and normally referred to as GOFAI: good old-fashioned artificial intelligence. As we will see in Section 2.1, this position is captured formally in the Physical Symbol Systems Hypothesis.

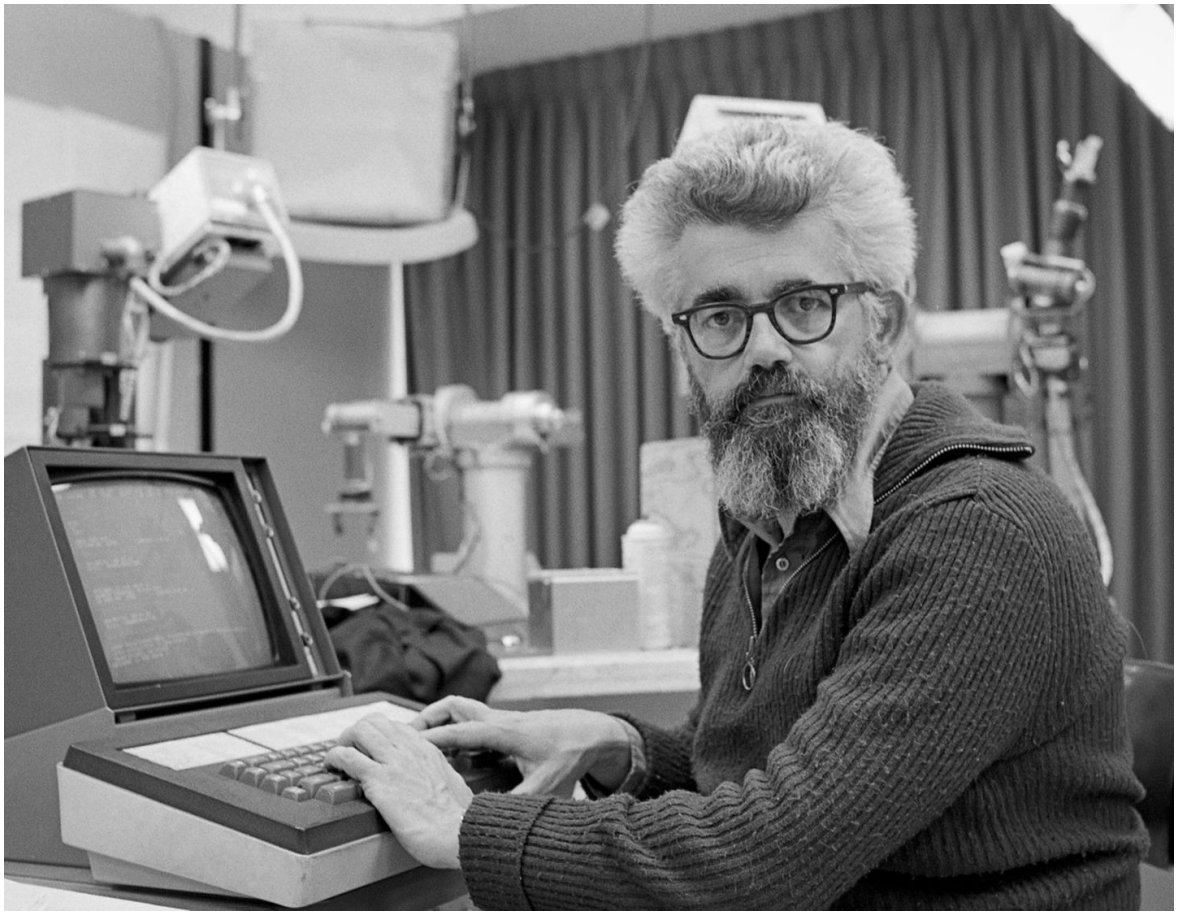

Fig. 1. John McCarthy, who coined the term artificial intelligence, at Stanford's Artificial Intelligence Laboratory in 1974.

However, AI has other roots in cybernetics, which is concerned with self-organization, regulation, and control (Wiener, 1948; Ashby, 1957). In 1950, Grey Walter developed two robotic turtles, Elmer and Elsie, that could roam around a room, find a charging station, and recharge themselves. In 1950, Claude Shannon demonstrated an electronic mouse, Theseus, that could navigate a maze. Both Walter's and Shannon's robots built on behaviorist psychology by using associative and reinforcement learning in relatively simple neural networks, rather than focusing on internal models and symbolic computation. Neural networks process information by propagating it through an interconnected layered network of relatively simple processing units: artificial neurons, very simplified versions of the neurons in biological brains. Loosely speaking, these artificial neurons weight and aggregate the information received and send a modified

version of the result to other processing units, typically in the next layer.[2] This approach, referred to as *connectionism*, progressed in parallel with the computationalist approach over the next sixty years and more. We'll say more about computationalist symbolic AI in Section 2.1 and connectionist AI in Section 2.2.

From the outset, symbolic AI was concerned with producing intelligent artifacts that exhibited the versatility, flexibility, and robustness of humans in rational problem solving. For this reason, it became known as strong AI. Despite the early optimism, strong AI proved to be very difficult to achieve. Consequently, AI techniques began to be applied in more limited domains with stronger constraints and a narrower focus. This approach became known as weak AI. Despite continual progress in both symbolic AI and connectionist AI in the 1970s and 1980s, performance on more challenging problems was disappointing and the popularity of AI waned during a period known as the AI winter. As we will see in Section 2.2, this was the second winter for connectionism. Despite the lack of success in applications, research proceeded apace in statistical techniques and on neural networks, just as it had done in the connectionist AI winter in the 1970s.

The AI winter came to an end in the 2000s when, building on research in the late 1990s, artificial neural networks with deeper network topologies, i.e. networks with many more layers than had been used in the mid-1980s to mid-1990s, and new learning techniques were introduced, leveraging the recent availability of much greater computing power in the form of graphic processor units (GPUs) and much larger datasets to train the networks. The period since 2011 has seen AI based on deep learning exhibit great success with many difficult applications in, for example, computer vision, robotics, and autonomous driving, natural language processing, sentiment analysis, medical imaging, and several other domains. This period also saw the development of some landmark probabilistic approaches to AI, perhaps the most celebrated of which is the Watson system from IBM (named after its founder Thomas J. Watson) and which won the TV show *Jeopardy!* in 2011, beating two human champions in answering rich natural language questions over a very broad domain of topics. The success of Watson was the result of probabilistic knowledge engineering that integrated many knowledge sources and exploited many techniques for search, hypothesis formulation, and hypothesis evaluation (Ferrucci et al., 2010).

AI had finally come of age, yielding reliable solutions to complex problems in many application domains. John McCarthy once remarked "As soon as it works, no one calls it AI anymore" (Meyer, 2011). This was no longer true.

At this point, many techniques which had traditionally not been part of AI, e.g. data science, optimization, and control theory began to be included under the purview of AI, contributing to its success. However, as Luc Steels points out, this expansion of AI departs from its original focus on strong AI which is human-centred, in the sense that the basis for its decisions is understandable by and communicable to humans, and human-level, in the sense that it displays the same versatility and flexibility that humans do (Steels, 2020). In contrast, the AI that derives from the behaviorist tradition, including connectionism, as well as the probabilistic data science models and statistical learning, are black box systems: the basis for their decision-making is not open to scrutiny by humans in any meaningful way, at least not at present. Consequently, much effort today is

[2] Artificial neural networks that propagate results to the next layer are called feed-forward networks, while networks that also propagate results back to previous layers are called recurrent neural networks. The simple two-layer networks developed in the 1960s are called perceptrons.

being expended to make this approach to AI more explainable and more trustworthy (XAI: eXplainable AI). This is not a trivial problem. While deep learning AI may achieve exceptional performance, by virtue of its statistical nature, it is susceptible to errors when faced with outliers and data that are not drawn from the distribution on which the system is trained. Such outliers may be very common, as is the case when systems are trained on data sets that exhibit implicit or explicit bias. The bias is modelled during training and the systems then inevitably operate in a biased manner, even if the bias in the original data set was not intended. We return to the issue of trustworthy AI in Section 4.3. For now, we take a closer look at the constituent approaches of AI: symbolic AI, connectionist AI, and statistical machine learning.

2. The Nature of AI

2.1 Symbolic AI and GOFAI

One of the key historical, methodological and epistemological approaches to AI is that of "Symbolic AI', often referred to as GOFAI (Good Old Fashion AI). This has its origins in the 1950s (i.e. part of the 1956 Dartmouth Workshop for the start of the AI movement) and constituted the primary, classical approach in the first 30 years of AI research, before the first AI Winter and the advent of Connectionist AI and machine learning (Boden 2014).

The term "symbolic" refers to the fact that AI algorithms and programs are based on a set of symbols and symbol manipulation processes. In fact, two of the founding fathers of symbolic AI, Allen Newell and Herbert Simon, proposed the concept of a **Physical Symbol System**, as a "a set of entities, called symbols, which are physical patterns that can occur as component of another type of entity called an expression (or symbol structure)" (Newell and Simon 1976: 116). These symbols are purely formal and meaningless entities, though in practice they are normally interpreted by the programmer with a particular semantic content such as words, numbers, pictures, actions etc. The symbolic expressions are created using logic formalisms, such as propositional logic with Boolean connectives (e.g. "Red AND Round") or predicate calculus (e.g. "Apple(Red, Round)"). They can also be arranged in IF-THEN production rules (e.g., "IF apple, THEN eat). In specific symbolic systems such as semantic networks, each node has a symbol ("Red", "Apple", "Fruit) with links having a label for the semantic relationship between node (e.g., "IS A" or "HAS") and hierarchical relationship between nodes. A collection of symbolic structures for a specific domain constitutes the knowledge base used by the system to reason about the problem. In general, symbols systems solve problems by using the processes of heuristic search (Newell and Simon 1976), where the search for the optimal link between the problem definition and its solution must be guided by heuristics, i.e., rules of thumb that are helpful in guiding the program toward the solution in an optimal way. The AI heuristic search and planning algorithms are widely used today for scheduling and logistics, for data mining, for games, for searching the web, and for planning in robotics.

An important aspect of the GOFAI approach is the idea that symbolic systems can model human intelligence. In fact, Newell and Simon (1976) proposed the **Physical Symbol Systems Hypothesis**, which states that "A Physical Symbols Systems has the necessary and sufficient means for general intelligent action" (1976:116). This is why

GOFAI systems have been applied to modelling mathematical reasoning, natural language processing, planning, game playing etc. A classic example of a GOFAI system is an expert system, i.e. a program that represents the knowledge of the human expert in a specific domain, using a a set of IF-THEN production rules, and which can be used to offer advice to non-experts or to provide solutions to experts. Beyond the historical examples of the first expert systems, such as Mycin to support medical doctors in the diagnosis and treatment of infectious diseases, nowadays expert systems have been developed in a wide range of domains (commercial, education, medical, and military applications), with some capable of highly complex planning on the order of tens of thousands of search steps (Franklin 2014).

The major strengths of GOFAI are its abilities to model hierarchical and sequential tasks, such as language processing, problem solving and games, and to represent knowledge bases using propositional contents and inference processes.

Some limitations of GOFAI systems are that these AI programs are brittle (i.e. that they can produce wrong and nonsensical decisions when there is missing or contradictory data), they are subject to the frame problem (i.e. the problem of representing what remains unchanged as a result of an action or an event) and the symbol grounding problem (i.e. linking symbols with the environment entities) and they cannot learn new knowledge. This, as well as the initial strong claims about the power of symbolic systems to deal with general intelligence and any problem domain, led to the first AI Winter in the 1980s, and the subsequent developments of connectionist and machine learning approaches (see next section). However, significant achievements of GOFAI includes the widespread use of commercial expert systems, their essential role in games industry (to control the intelligent behaviour of the virtual agents) including the historical victory of the IBM Deep Blue system in 1997 in beating the chess world champion Gasparov, and IBM Watson's victory in 2011 over two human champions in the *Jeopardy!* TV game (Franklin 2014).

2.2 Connectionist AI: From Perceptrons to Deep Neural Networks

Connectionist AI differs from symbolic AI in that information is processed by propagating it through an interconnected network of relatively simple processing elements, typically implemented as artificial neural networks. They use statistical properties rather than logical rules to analyze information. Although the term *connectionist model* is usually attributed to Feldman and Ballard (1982), the roots of connectionism reach back well before the computational era, with connectionist principles clearly evident in William James' nineteenth century model of associative memory (James, 1890).

Neural networks also have strong foundations in physics, as many of the mathematics concepts on neuron modelling and computation come from physics principles. For example, the Ising model (also known as the Ising-Lenz model), a mathematical model of ferromagnetism in statistical mechanics, provided inspiration for an model of associative memory (Little 1974) that was popularized by John Hopfield's recurrent neural network: the Hopfield net (Hopfield 1982). Boltzmann machines are variants of Hopfield nets that use stochastic rather than deterministic weight update procedures to avoid problems with the network becoming trapped in non-optimal local minima during training (Hinton and Sejnowski 1986). In the future, the principles of quantum mechanics may

provide the basis for efficient neural networks (Abbas et al. 2021), in particular, and for quantum AI (Dunjko Briegel 2018), in general.

The seminal paper by McCulloch and Pitts (1943),"A logical calculus immanent in nervous activity", is regarded as the foundation of artificial neural networks and connectionism (Anderson and Rosenfeld, 1988). Connectionism advanced significantly in the late 1950s with the introduction of the perceptron (Rosenblatt, 1958) and the Pandemonium model of learning (Selfridge, 1959), allowing artificial neural networks to be trained automatically instead of having to be tuned by hand. Network learning advanced further still in 1960 with the introduction of the delta rule for supervised training (Widrow and Hoff, 1960). However, perceptron networks suffered from a severe problem: no learning algorithm existed to allow the adjustment of the weights of the connections between input units and hidden units in networks with more than two layers, i.e., multi-layered perceptrons (MLPs). In 1969, Minsky and Papert (1969) showed that these perceptrons can only be trained to solve linearly separable problems and couldn't be trained to solve more general problems. This had a very negative influence on neural network research for over a decade. As a result, research on neural networks and connectionism suffered considerably and marked the beginning of a decade-long winter for connectionist AI (Pollack, 1989).

During the period that followed this disenchantment with perceptron networks, alternative connectionist models were developed, such as adaptive resonance theory (ART), first introduced by Stephen Grossberg in 1976 and developed in the ensuing years (Carpenter and Grossberg, 1995), and Teuvo Kohonen's self-organizing maps (Kohonen, 1982). ART addresses real-time supervised and unsupervised category learning, pattern classification, and prediction, while Kohonen networks exploit self-organization for unsupervised learning, and they can be used as either an auto-associative memory or a pattern classifier.

Perceptron-based neural networks underwent a strong resurgence in the mid-1980s with the introduction of the back-propagation algorithm (Rumelhart et al., 1986a,b), which had previously been derived independently by Paul Werbos (Werbos, 1974), among others (Medler, 1998). Backpropagation finally made it feasible to train MLPs, overcoming the restriction highlighted by Minsky and Papert (1969), thereby enabling MLPs to learn solutions to complex problems that are not linearly separable. This was a major breakthrough in neural network and connectionist research.

Perceptron-based neural networks typically represent a static mapping between the inputs and outputs in which data flows in just one direction through the network, from input to output. There is an alternative, however, in which the network has connections that loop back to form cycles, *i.e.* networks in which either the output or the hidden unit activation signals are fed back to the network as inputs. These are called recurrent neural networks. The recurrent pathways in the network introduce a dynamic behavior into the network operation.

By the early 2000s, the traditional neural network approach had fallen out of favor because effective training was limited to relatively small networks, both in terms of the number of layers and the number of units per layer, due to the lack of computational resources for training and the infeasible amount of time required to train large networks.

However, in the late 1990s, significant breakthroughs in deep networks, such as long short-term memory (LSTM) by Hochreiter and Schmidhuber (1997) and convolutional neural networks (CNNs) by LeCun et al. (1998), heralded a new era in connectionism,

although it took another ten years before they were widely adopted because of the lack of sufficiently large data sets and sufficient computational power for training. A CNN network is similar in principle to the multi-layer perceptrons of the 1980s and early 1990s, but they have more layers, each of which performs a different function. In a CNN, convolution refers to the application of a filter to the data being processed by the neural network. The key feature of a CNN is that these filters are learned by the network during the training phase. This marked a significant departure from previous approaches where the filters, and the features they extracted, were the result of hand-crafted design. Consequently, CNNs are able to map directly from the input space, e.g., the image to be classified or the image in which you want to search for a given object, directly to the image label or the object location. For this reason, they are referred to as end-to-end systems. The first CNN was created by Yan LeCun, focussing on handwritten character recognition (LeCun et al., 1998). In 2011, AlexNet (Krizhevsky et al., 2012), a CNN with seven hidden layers won the ImageNet Large Scale Visual Recognition Challenge.

Since then, deep neural networks have been applied successfully in many challenging applications (Schmidhuber, 2014; Goodfellow et al., 2016). The networks have become deeper, with twenty-two or many more layers, and performance has improved through the use of more effective activation functions (e.g. the rectified linear unit ReLU), the use of specialized layers (e.g. pooling), more advanced learning techniques (e.g. batch normalization and dropout), techniques to overcome the problem of vanishing gradients (where the error terms become too small to effect an improvement in network performance as they are propagated back in a deep network), and a better understanding of how to adjust the system hyper-parameters during training to improve performance.

While CNNs and regional CNNs (RCNNs) proved their mettle with very impressive performance in image recognition, object detection and localization, face detection, face recognition, and object tracking, new forms of recurrent neural networks proved very successful on problems that involve processing and analysing sequences of states, e.g. in natural language, by exploiting new recurrent elements such as long short term memory (LSTM) and gated recurrent units (GRU).

Progress has continued, with modern architectures successfully combining the power of deep CNNs and LSTMs to address problems that involve both images and language, e.g. automatic image annotation and captioning, image retrieval and synthesis base on linguistic descriptions (Mao et al., 2015).

The advent of generative adversarial networks, or GANs, which work as actor-critic systems, has provided the means for two learning networks to learn from each other and thereby improve the performance of both (Goodfellow et al., 2014; Goodfellow, 2017). This has yielded remarkable results in image synthesis, among many other applications.

Progress using deep neural networks for language understanding and generation has recently been advanced even further with the series of Generative Pre-trained Transformer (GPT) architectures, culminating, for now, in GPT-3 (Brown et al., 2020). Trained on trillions of words with some 175 billion machine learning parameters, GPT-3 is capable of generating natural language text that is often indistinguishable from that generated by humans.

2.3 Statistical and machine learning

A parallel development in AI in the last 20 years, with partial overlap with the AI connectionist approach, has been that of machine learning. This is the field primarily

based on a variety of **statistics-based inference methods** that use large data sets to estimate, i.e. learn, the parameters of a model with classification and predictive capabilities. This approach developed in conjunction with AI research in computer vision and speech (or more generally, pattern recognition), in robotics (e.g. reinforcement learning) and in neural networks (MLP and deep neural networks). Some people nowadays use the terms AI and machine learning interchangeably, especially because of the big, common emphasis on deep learning. But as we will see below, machine learning keeps a distinctive emphasis on data-driven statistical inference methods.

Amongst the various inferential strategies in statistics (e.g. analogical inference, domain-specific inference, and structural inference), the bulk of machine learning uses the structural inference approach. This uses domain-general algorithms which exploit the **internal structure of the data**, rather than identifying the semantic, domain-specific, content of the data. Structural inference is the basis of most machine learning frameworks, such as the well-known methods of regression, neural networks and Bayesian networks (Danks 2014). Given this data-centric (sometimes known as "data-hungry") approach, the recent, easy availability of potentially unlimited data from social media and the web, and wider access to cloud-based parallel computing systems such as GPUs (which are necessary to apply computationally-intensive statistical computations on large datasets) can in great part explain the recent, impressive contribution of machine learning to AI, and information technology in general. This is the case of the bootstrapping of neural network technology from the shallow MLP networks only trainable with small datasets in the 80s, 90s and early 2000s, to the deep CNN trained on huge datasets in the last 10 years.

Machine learning comprises a set of methods typically grouped into supervised and unsupervised techniques, as well as reinforcement methods. **Supervised learning** algorithms need a labelled dataset, i.e. where each data point (e.g. an image of a dog) is associated with a supervision signal or ground-truth (e.g. the category label "dog"). The learning algorithm has to find the parameters of the model (e.g. weights of a neural network) using the error between the model's own guess and the supervision label. Examples of supervised learning algorithms include MLP, CNN, and LSTM neural networks, decision trees, support vector machines and regression. **Reinforcement learning** can be considered part of the supervised approach, but where the supervision to learn a policy (e.g. actions that should be taken when certain sensory conditions prevail) is guided by a reward function (but see below for a view that separates supervised algorithms and reinforcement learning). **Unsupervised learning** algorithms do not require a labelled dataset, as they discover the regularity in the data and their organisation in separate categories. Example of unsupervised learning include the clustering algorithms such as k-means and autoencoder neural networks,

Yan LeCun, one of the founding fathers of deep learning, uses the metaphor of a cake to show how these methods are organised: "If intelligence is a cake, the bulk of the cake is unsupervised learning, the icing on the cake is supervised learning, and the cherry on the cake is reinforcement learning" (LeCun 2016). He has recently extended the concept of unsupervised learning, using the terms **self-supervised** and **predictive learning**. This is to refer to the power of unsupervised methods, such as autoencoders and word embeddings, that can automatically extract partial information from noisy or incomplete input data to predict the rest of the data.

An important set of machine learning approaches is that of Bayesian learning algorithms. The general Bayesian framework is based on the intuition that the beliefs after observing

some data is determined by the probability (prior probability distribution) of each possible explanation given that data. When processing a dataset, the machine learning algorithm uses the Bayesian rule to calculate the correct probability distribution over the hypotheses given that data. And given large datasets, the computations required for Bayesian learning become too difficult to be done analytically, thus the recent boost of Bayesian algorithms with easy access to parallel computational resources (Danks 2014).

Machine learning is to a great extent responsible for the recent successful developments in AI. The most successful and widely used applications in speech recognition, computer vision, natural language processing and robotics applications are based on deep learning and other Bayesian approaches. This includes the design of DeepMind's AlphaGo and AlphaZero systems, based on the combination of deep neural networks, reinforcement learning, and AI search algorithms, which, as we will see in Section 3.4 below, were able to outperform human champions in the Go game, even without human knowledge or supervision (Silver et al. 2016).

3. Example applications

3.1 AI applications in Medicine

The application of AI to medicine and healthcare has its origins in the early GOFAI developments of expert systems, such as the MYCIN for infectious diseases and DENDRAL on discovery of chemical compounds. More recently, the advent of machine learning and its focus on learning from data, has led to a resurgence of the development of medicine AI systems. These range from the use of deep learning for diagnosis of clinical images, on electronic health records, on medical sensors data and on the latest AI models on genomics and molecular and protein structure understanding (Miotto et al. 2018; Senior et al. 2020).

Deep learning methods, such as CNNs because of their impressive performance with 2D image recognition, have been widely used for **image-based cancer detection and diagnosis** (Hu et al. 2018). For example, in skin cancer diagnosis, the performance of CNNs to classify biopsy-proven clinical images (e.g. malignant melanomas versus benign nevi) was on par with that of 21 board-certified dermatologists. The AI system was trained using a dataset of 129,450 images with 2,032 different diseases (Esteva et al. 2017).

A recent landmark achievement in medicine and biochemistry is the **AlphaFold AI model for protein folding**. This was developed by Google DeepMind and was the winner in 2020 of the biennial Critical Assessment of protein Structure Prediction (CASP) competition. AlphaFold achieved a performance similar to the results from experimental methods. AlphaFold uses attention-based deep neural networks to interpret the structure of the spatial graphs, used to represent the proteins (Senior et al. 2020).

Medical AI applications present significant **technological and ethical challenges**. One key issue is the reliance on the quality and variety of the training data, as healthcare datasets typically are sparse, noisy, heterogeneous, and time-dependent. Moreover, new methods and tools are needed to enable interactive machine learning to interface with healthcare information workflows, keeping the human in the loop (Miotto et al. 2018; Holzinger 2016). There are also important ethical considerations. One is for example the need for explainable systems so that clinicians (both novice and expert doctors) can

access causal explanations of the AI's decision-making process (Holzinger et al. 2019). We return to this issue of trust and explainability in AI in Section 4.3.

3.2 AI applications in Robotics

Robots feature prominently in the general public's perception of artificial intelligence. While there is a long way to go to achieve what they are capable of in science fiction movies, impressive progress in mechatronics and control has been achieved over the past ten years, for example in the mobility displayed by robots such as Atlas (Atlas, 2021) from Boston Dynamics (Boston Dynamics, 2021) and the dexterity of the Shadow Hand (Shadow Hand, 2021). However, despite recent advances in cognition-enabled manipulation in everyday activities (EASE, 2021), the same cannot be said for the cognitive capabilities of robots. Nevertheless, robotics remains a key element of the domain of AI, and AI tools and techniques play a central role in achieving the robust performance that are required of robots, especially when they are operating in environments that are not engineered to facilitate their operation.

AI is used in robotics for many purposes, including autonomous navigation, task planning, task execution, object detection, object grasping and manipulation, inspection & surveillance, social human-robot interaction, including natural language processing, facial recognition, sentiment analysis, gesture understanding, and intention recognition. It is also used in an extensive range of robots. At the time of writing, the IEEE robots website (IEEE Robots, 2021) features 229 robots of many different types: wheeled, legged, tracked, airborne, underwater, and humanoid, targeting consumers, entertainment, education, research, medicine and health care, disaster response, service & industrial, aerospace, military & security applications, telepresence, self-driving cars, and agriculture. Perhaps the epitome of AI in robotics is the goal of creating a collaborative robot, *i.e.* one that can share a common goal and share the human's intentions to achieve that goal, acting jointly with the human, paying attention to what the human is doing, and, crucially, anticipating any help the human might need to complete whatever tasks she or he is working on.

The **overlap between robotics and AI** is a good illustration of the manner in which the scope of AI has expanded to embrace many techniques in computer science and engineering, e.g., control theory, machine learning, signal and image processing, as we noted in the introduction. In addition, AI techniques are used to support the core cognitive abilities of perception (i.e., the interpretation of sensed data), attention, action selection, memory, learning, reasoning, metacognition, and prospection (Kotseruba and Tsotsos, 2020).

One example of AI in robotics is robot-enhanced therapy (Cao et al., 2019) where **robots assist a psychotherapist working with children** with autism spectrum disorder, specifically next generation robot-enhanced therapy. Under the guidance of clinical practitioners, this project developed interactive capabilities for social robots that allowed them to engage a child in clinically derived exercises. The robot can operate autonomously for limited periods under the supervision of a psychotherapist. AI plays a major part in the success of this application, specifically in its cognitive ability to interpret body movement and appearance-based cues of emotion. This allows the robot to assess the child's actions by learning to map them to therapist-specific classes of behavior. In turn, the robot also learned to map these child behaviours to appropriate robot responses, again as specified by the therapists.

One of the goals of cognitive robotics is for humans to be able to give a robot some task to do by stating that task in the same terms they would use if they were talking to another human being, conveying just the essence of the goal without have to specify exactly how that task is the be carried out. Fig 2. shows a PR2 robot in the process of pouring popcorn from a saucepan during a demonstration of cognitively-enabled robot manipulation using the CRAM cognitive architecture (Beetz et al. 2010).

Fig. 2. A PR2 robot demonstrating cognition-enabled manipulation by using knowledge and reasoning to determine the motions required to pour popcorn from a saucepan (Sandini et al. 2021).

3.3 AI applications in the Web and Social Media

AI is having a tremendous impact in a variety of applications and functionalities for the web (e.g. search algorithms, music and video recommendations, automatic translation) and social media (e.g. news selection and recommendation, sentiment analysis, face recognition). Although this progress is resulting in clear benefits to people and society, it also carries important ethical considerations and risks.

AI has significantly changed the **search algorithms** for the web. For example, Google's initial PageRank algorithm (based on standard mathematical methods) has now developed in a collection of search tools, such as the Hummingbird framework. This, for example, complements PageRank's results with RankBrain, based on machine learning algorithms for entity recognition, and the recent introduction of BERT (Bidirectional Encoder Representations from Transformer), which uses a neural network for natural language processing. BERT uses word context to find more semantically-relevant information, allowing it to select 'featured snippets', i.e., short snippets of text, figures, or tables that appear at the top of Google's search results and provide definitions of the searched-for item. In fact, Google has become a "machine learning first" company (Levy 2016).

Another example of the widespread use of AI and machine learning algorithms in the web is for **recommender systems**. This, for example, concerns recommendation for related purchases in e-commerce sites, suggestions of related news and friends in social media, and personalised recommendation in media streaming sites and apps. 80% of movies watched on Netflix are based on AI recommendations (Zhang et al. 2019). As in other domains, deep machine learning systems have become the default algorithm for the latest recommendation systems. This raises important issues for consideration, such as the need for machine reasoning and explainable recommendation with back-box deep neural, and the ethical implications concerning the influence of these algorithms in politics (e.g. introducing bias into recommendations and social media content during elections), in public health (e.g. undermining scientifically-grounded health advice) and in the generation and diffusion of fake information, generally.

AI applications for **face recognition** have also become widespread in the web and in social media. These algorithms can be used for image matching and people recognition (e.g. in social media photo tagging) as well as for authentication (e.g. to implement secure access in some smartphone systems). A variety of AI and machine learning algorithms have been developed to implement this functionality (e.g. Bayesian and support vector machines) with the recent design of a variety of deep learning face recognition systems typically based on CNN and autoencoders (Guo & Zhang 2019). However, face recognition algorithms based on learning from datasets have important ethical implications, e.g. regarding possible biases in the data used for the training. For example, In 2018, a seminal paper by computer scientists Buolamwini and Gebru (2018) demonstrated that leading facial recognition systems produced substantial disparities in the accuracy of gender classification, e.g. with error rates of up to 34.7% in the classification of darker-skinned females (whilst the maximum error rate for lighter-skinned males was 0.8%). This highlights the urgent need to address and remedy implicit bias in such systems and make sure they are based on fair, transparent, and accountable facial analysis algorithms.

3.4 AI applications in Sports

While the use of statistical analysis is well-established in sport, AI is taking it to a new level. Possible applications range across the entire spectrum of activities. Barlow and Sriskandarajah (2019) identify eleven applications across seventeen sports which are being or will be impacted by AI. These applications include identifying talent and determining optical game strategies. Together with wearable sensors, AI can also assist with training by providing advice on optimal nutrition, enhancement of physical skills, and recovery management, much like recommender systems do for online marketing and decision support in business.

AI technologies such as computer vision have been used routinely to **assist with umpiring during games**, especially using automated ball tracking and line calling applications. For example, the Hawk-Eye system (Hawk-Eye, 2021) visually tracks the trajectory of the ball using six high-speed cameras, the images from which are used to triangulate the ball's position over time, displaying a virtual reality trajectory of its statistically most likely path. While the system is widely used in many sports and is accurate to within 3.6 mm, there is also some debate about whether the system's decisions should be accompanied by a confidence value. In this context, Collins and Evans (2008) note that systems such as Hawk-Eye illustrate the difference between the

models of the world that AI uses to make decisions and the uncertain reality of the world they model. Such systems are also used to provide statistical information on the performance of players when training and on competitors when preparing for matches.

AI can also be used for **automated generation of video highlights**, while integrated vision and natural language technology can be used for automated generation of copy for publication in print and online. Chatbots are used to enhance the experience of fans and also support interaction with the media. Drone cameras are used to provide game footage and enhance security. Other examples include optimisation of policies for stadium entry and the use of smart ticketing.

The All England Lawn Tennis Club hosts the annual tennis championship at Wimbledon and uses IBM's Watson technology to provide a variety of services, including **real time match reports and uncovering player insights**. It also powers a voice-activated cognitive assistant "Fred", named after the late champion Fred Perry, to help spectators find their way around the venue (Shaw, 2017). In 2017, in order to determine what attributes make a great champion, IBM Watson analyzed tennis champions across six broad categories including passion, performance under pressure, serve effectiveness, stamina, how well the player either adapted their normal playing style to an opponent or was able to force an opponent to conform to their tactics, and the ability to return serves. It used 22 years of unstructured data and analyzed an estimated 53,713,514 tennis data points, captured since 1990. Watson's Personality Insights API helped uncover player traits and behaviors based on previous player interviews, information which served to engage and inspire discussion among experts, sports commentators, and fans. For example, Watson's analysis revealed that a player's serve had improved every year and was the driving force to her becoming a Wimbledon champion (Shaw, 2017).

To identifying successful game strategies, an **AI system can play against itself**, as the DeepMind AlphaGo system did, before going on to beat Lee Sedol, the winner of 18 world titles, in 2016, and achieve 60 straight wins in time-control games against top international players in 2017 (AlphaGo, 2021). The original version of AlphaGo used two neural networks, a policy network that produces moves and a value network that evaluates board positions. The policy network was trained by supervised learning based on human expert moves and subsequently refined by reinforcement learning by playing against itself. Subsequently, in AlphaGo Zero, even better performance was achieved based purely on reinforcement learning without any prior supervised training. Apart from its formidable performance, what is significant about AlphaGo is that it uncovered several innovative strategies that greatly surprised expert players, demonstrating the potential for AI to augment human abilities and exceed human performance.

4. Future challenges

4.1 Collaborating with machines and robots

AI has contributed significantly to the design of intelligent control models and cognitive architectures for sensorimotor behaviour (e.g. perception, navigation manipulation) and cognitive capabilities (e.g. planning, language) in robots, as we have seen in section 3.1. But major challenges still remain for the realisation of above skills, specifically to allow the robot to handle the complexity of real-word scenarios, i.e. cluttered, dynamic, unpredictable environment where objects to be grasped or obstacles to be avoided are

difficult to see, can be occluded, or change their position over time. Beyond the complexity of designing skills in individual robots, a significant challenge for robots, and intelligent machines in general, is that of handling **interaction with people for collaborative tasks**, also known as human-robot interaction (HRI) and social robotics (Bartneck et al. 2020). This type of interaction includes a variety of scenarios, such as joint action in flexible manufacturing setup between a worker and a cobot (collaborative robot), assistive robot companions for older and disabled people or in hospital and care homes, and robot tutors for education or entertainment. Within the field of HRI and collaborative human-machine interaction, some of the challenges for future research on the combination of AI and collaborative robots concern (i) the skills required interaction, such as the ability to read a human's intentions and form a theory of mind, i.e. modelling her or his goals and state of mind, (ii) the mode of interaction with machines with variable autonomy, and (iii) the quality of interaction for long-term, trustworthy interaction that fosters well-being.

Fig. 2. This sequence of pictures depicts a situation in which the iCub humanoid robot (www.icub.org) is interacting with a human, reading her intention to get her phone from her bag, and alerting her to the fact that it is on the desk, hidden from her by the laptop. This sequence has been staged to illustrate the future capabilities of a cognitive robot and has not yet been implemented (Sandini et al. 2021).

The research challenge on the use of AI for the design of the social and cognitive skills for interaction include for example the capability of **intention reading** and the implementation of an artificial **Theory of Mind** (Vinanzi et al. 2019). Intention reading is the capability of the robot to detect the human user's intended goal of the joint interaction. For example, when a cobot is working with a person to assemble a table, it must

anticipate the goal and the next action that the person is expecting the robot to perform. Theory of Mind describes a more general view of intention reading, as this concerns the robot's capability to understand and predict the belief, desires and goals of the person; see Fig. 3. AI methods, such as Bayesian networks and deep learning, can be used to build artificial Theory of Mind skills in robots (Vinanzi et al. 2019).

The mode of interaction between people and robots and machines concerns the concept of **variable autonomy**. Rather than considering the extremes of a fully autonomous machine (e.g. autonomous car without a driver) or of a fully tele-controlled robot (e.g. remote control of a mobile robot in a nuclear site), most of tasks in which intelligent robots will interact with people concern a variable degree of collaboration, requiring a robot with a variable degree of autonomy to adapt to the user needs and to the environmental circumstances. This could be for example the case of intelligent assisted driving and the six levels of driving automation identified by the car industry. Future intelligent vehicles will dynamically switch from situations in which the car performs some lane-following, steering and acceleration tasks autonomously, when it can easily recognise the road and traffic condition, whilst the driver must take control when the car is unable to perform the driving task (level 3 of conditional driving automation).

Another important future research direction in AI for collaborative robotics concerns the quality of the interaction, i.e. the design of **long-term and trustworthy interaction and wellbeing in human-robot collaboration**. Long-term interaction requires the robot to be able to engage in continuous, meaningful, and contextualised interaction over a series of interactions lasting for days, weeks, or longer. This will require the ability to recognize the person and their personality and preferences, to remember recent interactions, and to engage in empathic behaviour with the person needs (Leite et al. 2013). This is a significant challenge in robotics, as the great majority of current intelligent robots are only capable of short term (typically one-session only) interaction. Trustworthy interaction, a growing field of research, requires people's acceptance and trust of the robot's behavior and decision making process. This is also linked to ethical issue explainable AI (see section 4.4) and to the achievement of peoples' and robot's reciprocal theory of mind (Mou et al. 2020; Vinanzi et al. 2021).

4.2 Self-learning and self-programming machines

The quest for the automatic generation of computer programs, also known as program synthesis or self-programming machines, has been one of the main challenges of AI since the outset. Since the first symbolic (GOFAI) approaches to AI, and systems such as the General Problem Solver (Newell and Simon 1976), the idea was that AI systems can use general-purpose knowledge to generate new text, solve mathematical and practical problems, and create new computer programs. In addition, with the advent of machine learning approaches, AI has started to put emphasis on self-learning machines, which can learn with no or minimal supervision from humans.

This **self-programming machine** challenge has only recently received a significant boost through the combination of deep learning and NLP methods. For example, DeepCode is a code generator that uses a neural network to predict the properties of the program that can produce the outputs given specific the inputs (Balog et al. 2016). SketchAdapt, i.e. a system that learns, without direct supervision, when to rely dynamically on pattern recognition and when to perform symbolic search for explicit reasoning (Hewitt and Tenenbaum 2020). This mimics the human ability to dynamically

incorporate pattern recognition from examples and reasoning to solve programming problems from examples or natural language specification.

Recently, the GPT-3 system (Brown et al. 2020) has been proposed for natural language generation, with potential application to automatic program synthesis. GPT-3 is a large scale deep learning model for natural language processing with an order of magnitude more parameters than any previous NLP model. The model can generate new text without the need of further training or task-specific fine tuning of its parameters. This has been evaluated with few-shot demonstrations, i.e. via text interaction with the model giving the task description and one or few examples. This model demonstrates strong performance on many NLP datasets on translation and question-answering. It can also perform several tasks that require on-the-fly reasoning, such as unscrambling words, using a novel word in a sentence, or performing arithmetic. GPT-3 can produce samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. In addition, GPT-3 has been used for generating programs, such as for the code to create the Google home page (Heaven 2000).

Regarding the challenge of creating a **self-learning machine**, the first attempts to design AI systems and robots that autonomously learn without supervision from humans have recently been realised in developmental robotics (also known as autonomous mental development). This area of robotics takes inspiration from child development to design robots that go through stages of developmental for the incremental acquisition of sensorimotor and cognitive skills (Cangelosi & Schlesinger 2015). Developmental robots use intrinsic motivation mechanisms (e.g. implemented with reinforcement learning) to allow them to initiate and manage self-learning via curiosity-driven mechanisms for an open-ended, cumulative acquisition of skills.

Another example of self-learning AI is the AlphaZero system mentioned above in which artificial agents play the game Go against each other, bootstrapping their final learning capabilities. This led to the acquisition of skills that far outperformed the skills of the best human players (Silver et al. 2017; see also section 2.3).

Returning to LeCun's cake metaphor for AI and machine learning (section 2.3), his suggestion that the cherry on the cake is reinforcement learning has been revised and widened, with the cherry now being predictive, self-supervised learning. Here, the agent generates its own labels and teaching input, e.g. with autoencoders or Word2vec unsupervised learning methods, so that the system predicts the output from partial, incomplete, or self-generated input. This suggests that AI will increasingly be base on predictive, self-learning methodologies (Peng 2019).

4.3 Social and Ethical Aspects of AI

AI can bring significant benefits to all. However, the examples we have given so far focussed on applications in the developed world and, indeed most of the national strategies on AI have been created by governments in developed countries (OEC- DAI, 2021). Nevertheless, the fourth industrial revolution in general, and AI in particular, are just as relevant for developing countries. For example, AI is having an increasingly positive impact in Africa, in sectors such as energy, healthcare, agriculture, public services, and financial services (Novitske, 2018, Alupo et al., in press). It has the potential to drive economic growth, development, and democratization, to reduce poverty, to improve education, to support health-care delivery, to increase food production, to

improve the capacity of existing road infrastructure by increasing traffic flow, to improve public services, and to improve the quality of life of people with disabilities (Pillay and Access Partnership, 2018).

Paradoxically, the deployment of AI in developed countries can have a severe negative impact on developing countries due to the phenomenon known as premature deindustrialization (Rodrik, 2016; Kozul-Wright, 2016) which sees low-wage developing countries having fewer opportunities for industrialization before achieving income levels comparable to those in developed countries. Developing countries lose their competitive advantage in manufacturing due to the lower cost automation in developed countries and therefore miss out on the economic benefits that developed countries enjoyed as their workforces moved from low-value work to manufacturing before progressing to a post-industrial service economy. Consequently, developing countries are increasingly likely not to have the opportunity for rapid economic growth by shifting workers from farms to factory jobs because (a) automation undermines the labor cost advantage and (b) developments in robotics and additive manufacturing allow companies in advanced economies to locate production closer to domestic markets in automated factories, allowing this work to be moved closer to home in the developed countries.

AI can also be used for negative purposes, either intentionally or unintentionally, e.g., by fomenting religious, ethnic, social, and political divisions through fake misinformation created by deep networks (Besaw and Filitz, 2019). Of particular concern is the issue of implicit and explicit bias in the data that are used to train the AI models, thereby resulting in discrimination against people on the basis of gender or race. Examples of bias against dark-skinned people include in facial analysis (Buolamwini and Gebru, 2018), pedestrian detection (Wilson et al., 2019), and in predicting recidivism (Larson et al., 2016).

There is also the issue of democratization in AI, i.e., open access to AI technology by developers everywhere, in both developing and developed countries, Training deep neural networks requires access to large expensive computational resources which may be out of reach of some. Training also requires very large data sets and these may not be available. For example, in efforts to use machine learning to make the web available in local African languages progress is being inhibited for so-called “low- resourced” languages, i.e. languages for which few digital or computational data resources exist (Nekoto et al., 2020), because of the lack of sufficient training data. This is symptomatic of a problem that is endemic to almost all applications of machine learning in developing countries: the paucity of data. It is crucially important that the fourth industrial revolution, powered by AI, happens in a fair, ethical manner (EUAI, 2021).

4.4 Intelligence, Brains, and Consciousness

Why does intelligence matter? Indeed, what is intelligence? You should be surprised by the fact that we haven’t yet answered this question. We avoided it by defining AI as the endeavour to mimic and augment human intelligence and we avoided saying what human intelligence is. This has not caused us any problems because we all take human intelligence for granted. But let’s pause here to consider what we mean by intelligence and why it is so important to be intelligent. The answers will reveal why we have brains — the seat of intelligence — and what brains do. This will lead to other questions that are harder to answer, such as how consciousness fits into the picture.

Let's start by answering the question: what is intelligence? There are many possible answers but the one that has the most appeal derives from the answer to a different question: why do we have brains? The neuroscientist Daniel Wolpert provides an unexpected but compelling answer. He argues that we have brains to allow us to control movement (Wolpert, 2011). This mirrors what Francisco Varela and Umberto Maturana say about cognition: "Cognition is effective action" (Maturana and Varela, 1987). From this perspective, we see intelligence as the way to be effective in our control of our movements and in the way we act in the world. The key to understanding why this is so important — and so difficult — is to see that the number of possible ways we can move and act, and the number of possible outcomes of these movements and actions, is infeasibly large if we are to consider all the possibilities and choose the best one, or even a good one. This is what Allen Newell and Herbert Simon pointed out in their Turing Award (a sort of Nobel Prize for computer scientists) lecture: "The task of intelligence, then, is to avert the ever-present threat of the exponential explosion of search" (Newell and Simon, 1976), i.e. the search for good ways to act. Newell and Simon were referring to the search for the solution to a problem, but it amounts to the same thing. This is a satisfyingly straightforward and very practical way of understanding intelligence and the brains that give rise to intelligence.

However, brains are even better than that. They also predict the need to act and the outcome of those actions, and they do so all the time, at every instant, as we act and as we anticipate the future, milliseconds ahead, seconds ahead, hours, days, years. Indeed, it has been argued that brains are, in effect, probabilistic (meaning they can deal effectively with uncertainty) prediction machines (Friston, 2010; Downing, 2009; Seth, 2015).

But what then of consciousness? Can intelligent machines also be conscious? Can we take AI even further and build machines with artificial consciousness. Many people think this is a distinct possibility. Indeed, according to Paul Vershure, "understanding the nature of consciousness is one of the grand outstanding scientific challenges" and he proposes a scientifically-grounded approach to addressing the challenge of answering the question of what consciousness is and how physical systems can give rise to it (Verschure, 2016). He maintains that this challenge stands at the centre of knowing what it is to be human. In this, he reprises the original motivation for studying AI and one which comes full circle in the light of the great advances in AI in the past sixty-five years and the manner in which, as we said at the outset, AI affects the way humans live, work, and relate to one another.

5. Summary and Conclusion

Artificial intelligence impacts on all aspects of human activity: it automates tasks, it assists with decision-making, it augments and extends our cognitive capabilities, and it can even operate autonomously, if we allow it, without recourse to human oversight.

AI began as an attempt to understand and replicate human intelligence, initially taking two routes to that goal, one via connectionism, and one via symbolic computationalism, reflecting their inspiration by behaviorist and constructivist psychology, respectively. These two approaches waxed and waned in their own respective ways over the decades, to be joined in the 1980s by machine learning and in the 1990s by statistical machine learning, probabilistic inference networks, and other established disciplines in computer science. Breakthroughs in deep neural network learning and deep neural network

topologies, aided by very large data sets and equally large increases in processing power, yielded great success in many application domains. The symbolic knowledge representation and reasoning approach also developed rapidly, especially in cognitive architectures, as knowledge bases and ontologies increased greatly in size and sophistication and as the hybrid paradigm, combining symbolic approaches and sub-symbolic connectionist approaches were developed, e.g. in cognitive architectures such as Soar (Laird, 2012), ACT-R (Anderson et al., 2004), CLARION (Sun, 2016), among others.

While the success of statistical machine learning in narrow targeted applications yielded great success, it did so at the expense of losing focus on AI's original goal of understanding and replicating human-level intelligence. There has been a resurgence of interest in what is now known as Artificial General Intelligence (AGI) in cognitive science and cognitive systems. Still, the ultimate goal of replicating the versatility of human cognition remains elusive and it is unclear when it will be achieved. What is certain is that the AI quest will continue and AI in its many guises will continue to permeate our lives, change them, hopefully for the better.

In seeking to steer the path to the future, it is likely that other strands of thinking will woven into the fabric of AI, especially concerning the trustworthiness of AI in autonomous systems, i.e. its role in serving the bigger agenda of creating self-maintaining systems that can operate robustly and prospectively in the face of uncertainty and that can continually develop through self-programming as they interact with and learn from the world and the people in it. While there is much important work yet to be done to promote the development of democratized, trustworthy, ethical AI in the developed and developing worlds, an equal challenge will be how to control the role of AI in autonomous systems, possibly conscious ones, where the relationship with humans is no longer symbiotic. We are far from that point at present but it is likely we will reach it, and everything will change quickly when we do.

In Ernest Hemingway's novel *The Sun Also Rises* there is a dialog between two characters which goes as follows. "How did you go bankrupt?" Bill asked. "Two ways," Mike said. "Gradually and then suddenly." And so it will be with autonomous AI. Our collective responsibility is to work together in a directed manner during the present gradual phase so that, when the full impact of AI is suddenly felt, it will be for the greater good of all humankind.

Acknowledgements

In the preparation of this article, Angelo Cangelosi's work was partially supported by the Air Force Office of Scientific Research, USAF under Award No. FA9550-19-1-7002, by the UKRI TAS Node on Trust (EP/V026682/1) and the H2020 projects PERSEO, TRAINCREASE and eLADDA.

References

AlphaGo (2021) URL https://deepmind.com/research/case-studies/alphago-the-story-so-far

Alupo CD, Omeiza D, Vernon D (in press) Realizing the potential of ai in africa: It all turns on trust. In: Ferreira MIA (ed) Towards Trustworthy Artificial Intelligence Systems, Springer

Anderson JA, Rosenfeld E (eds) (1988) Neurocomputing: Foundations of Research. MIT Press, Cambridge, MA

Anderson JR, Bothell D, Byrne MD, Douglass S, Lebiere C, Qin Y (2004) An integrated theory of the mind. Psychological Review 111(4):1036–1060

Ashby WR (1957) An Introduction to Cybernetics. Chapman and Hall, London

Atlas (2021) Atlas humanoid robot by boston dynamics

Balog, M., Gaunt, A. L., Brockschmidt, M., Nowozin, S., & Tarlow, D. (2016). DeepCoder: Learning to write programs. arXiv preprint arXiv:1611.01989.

Barlow A, Sriskandarajah S (2019) Artificial Intelligence – application to the sports industry. URL https://www.pwc.com.au/industry/sports/artificial-intelligence-application-to-thesports-industry.pdf

Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M., & Šabanović, S. (2020). *Human-robot interaction: An introduction*. Cambridge University Press.

Besaw C, Filitz J (2019) AI & global governance: AI in Africa is a double-edged sword. Tech. rep., United Nations University, URL https://cpr.unu.edu/ai-in-africa-is-a-doubleedged-sword.html

Boden M. (2014). GOFAI. In The Cambridge handbook of artificial intelligence Frankish, Keith, editor.; Ramsey, William M., 1960- editor.

Boston Dynamics (2021) URL https://www.bostondynamics.com/

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners. arXiv:2005.14165.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). Proceedings of Machine Learning Research 81:1–15.

Cangelosi, A., & Schlesinger, M. (2015). Developmental Robotics: From Babies to Robots. MIT press.

Cao H, Esteban PG, Baxter P, Belpaeme T, Billing E, Cai H, Coeckelbergh M, Costescu C, David D, Beir AD, Garc´ıa DH, Kennedy J, Liu H, Matu S, Mazel A, Pandey AK, Richardson K, Senft E, Thill S, de Perre GV, Vanderborght B, Vernon D, Yu KWH, Zhou X, Ziemke T (2019) Robot-enhanced therapy: Development and validation of a supervised autonomous robotic system for autism spectrum disorders therapy. IEEE Robotics and Automation Magazine 26(2):49–58

Carpenter GA, Grossberg S (1995) Adaptive resonance theory (ART). In: Arbib MA (ed) The Handbook of Brain Theory and Neural Networks, MIT Press, Cambridge, MA, pp 79–82

Collins H, Evans R (2008) You cannot be serious! public understanding of technology with special reference to “hawk-eye”. Public Understanding of Science 17(3):283–308
Danks D. (2014). Learning. In The Cambridge handbook of artificial intelligence Frankish, Keith, editor.; Ramsey, William M., 1960- editor.
Downing K (2009) Predictive models in the brain. Connection Science 21:39–74
DREAM (2020) Development of robot-enhanced therapy for children with autism spectrum disorders. URL https://dream2020.github.io/DREAM/
EASE (2021) Everyday activity science & engineering. URL https://ease-crc.org/
Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–18.
EUAI (2021) European Commission ethics guidelines for trustworthy AI. URL https://ec.europa.eu/futurium/en/ai-alliance-consultation
Feldman RA, Ballard DH (1982) Connectionist models and their properties. Cognitive Science 6:205–254
Ferrucci D, Brown E, Chu-Carroll J, Fan J, Gondek D, Kalyanpur AA, Lally A, Murdock JW, Nyberg E, Prager J, Schlaefer N, Welty C (2010) Building watson: An overview of the deepqa project. AI Magazine 31(3):59–79
Franklin S. (2014). History, motivations, and core themes. In The Cambridge handbook of artificial intelligence Frankish, Keith, editor.; Ramsey, William M., 1960- editor
Frey C, Osborne M, Holmes C, Rahbari E, Curmi E, Garlick R, Chua J, Friedlander G, Chalif P, McDonald G, Wilkie M (2016) Technology at work v2.0: The future is not what it used to be. White paper, Oxford Martin School, University of Oxford, and Citigroup
Friston KJ (2010) The free-energy principle: A unified brain theory? Nature Reviews Neuroscience 11(2):127–138
Goodfellow I (2017) NIPS 2016 tutorial: Generative adversarial networks. ArXiv preprint ArXiv:1701.00160v4
Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in Neural Information Processing Systems 27, Curran Associates, Inc., pp 2672–2680
Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press
Guo, G., & Zhang, N. (2019). A survey on deep learning based face recognition. Computer vision and image understanding, 189, 102805.
Hawk-Eye (2021) URL https://www.hawkeyeinnovations.com/
Heaven, W. D. (2020). OpenAI’s new language generator GPT-3 is shockingly good—and completely mindless. MIT Technology Review.
Hewitt, L. B., Le, T. A., & Tenenbaum, J. B. (2020). Learning to infer program sketches. In Proceedings of the 36th Conference Conference on Uncertainty in Artificial Intelligence.
Hochreiter S, Schmidhuber J (1997) Long short-term memory. In: Neural Computation, vol 9, pp 1735–1780
Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop?. Brain Informatics, 3(2), 119-131.
Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.

Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., & Sun, Q. (2018). Deep learning for image-based cancer detection and diagnosis− A survey. Pattern Recognition, 83, 134-149.

IEEE Robots (2021) URL https://robots.ieee.org/

James W (1890) The Principles of Psychology, vol 1. Harvard University Press, Cambridge, MA

Kelly JE (2015) Computing, cognition and the future of knowing. White paper, IBM Corporation

Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biological Cybernetics 43:59–69

Kotseruba I, Tsotsos J (2020) 40 years of cognitive architectures: core cognitive abilities and practical applications. Artificial Intelligence Review 53(1):17 – 94

Kozul-Wright R (2016) Robots and industrialization in developing countries. Tech. rep., United Nations Conference on Trade and Development (UNCTAD)

Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in Neural Information Processing Systems, vol 25

Laird JE (2012) The Soar Cognitive Architecture. MIT Press, Cambridge, MA

Larson J, Mattu S, Kirchner L, Angwin J (2016) How we analyzed the COMPAS recidivism algorithm. ProPublica (5 2016) 9(1)

LeCun Y. (2016). NIPS2016 Invite Talk https://www.youtube.com/watch?v=Ount2Y4qxQo&t=1072s

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324

Leite, I., Martinho, C., & Paiva, A. (2013). Social robots for long-term interaction: a survey. *International Journal of Social Robotics*, *5*(2), 291-308.

Levy, Steven. "How Google is Remaking Itself as a Machine Learning First Company." Wired-Backchannel, June 22 (2016).

Licklider JCR (1960) Man-Computer Symbiosis. IRE Transactions on Human Factors in Electronics HFE-1:4–11

Manyika J, Lund S, Chui M, Bughin J, Woetzel J, Batra P, Ko R, Sanghvi S (2017) Jobs lost, jobs gained: Workforce transitions in a time of automation. Tech. rep., McKinsey Global Institute

Mao J, Xu W, and Z Huang YY, Yuille A (2015) Deep captioning with multimodal recurrent neural networks (m-RNN). In: Proceedings of the 3rd International Conference on Learning Representations (ICLR)

Maturana H, Varela F (1987) The Tree of Knowledge — The Biological Roots of Human Understanding. New Science Library, Boston & London

McCulloch WS, Pitts W (1943) A logical calculus of ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5:115–133

Medler DA (1998) A brief history of connectionism. Neural Computing Surveys 1:61–101

Meyer B (2011) John mccarthy. URL https://cacm.acm.org/blogs/blog-cacm/138907-johnmccarthy/fulltext

Minsky M, Papert S (1969) Perceptrons: An Introduction to Computational Geometry. MIT Press, Cambridge, MA

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236-1246.

Mou W., Ruocco M., Zanatto D., Cangelosi A. (2020). When would you trust a robot? A study on trust and theory of mind in human-robot interactions. *Proceedings of RO-MAN2020, 29th IEEE International Conference on Robot and Human Interactive Communication*, Naples, August 2020

Nekoto W, Marivate V, Matsila T, Fasubaa T, Fagbohungbe T, Akinola SO, Muhammad S,
Kabongo Kabenamualu S, Osei S, Sackey F, Niyongabo RA, Macharm R, Ogayo P, Ahia O, Berhe MM, Adeyemi M, Mokgesi-Selinga M, Okegbemi L, Martinus L, Tajudeen K, Degila K, Ogueji K, Siminyu K, Kreutzer J, Webster J, Ali JT, Abbott J, Orife I, Ezeani I, Dangana IA, Kamper H, Elsahar H, Duru G, Kioko G, Espoir M, van Biljon E, Whitenack D, Onyefuluchi C, Emezue CC, Dossou BFP, Sibanda B, Bassey B, Olabiyi A, Ramkilowan A, Oktem A, Akinfaderin A, Bashir A (2020) Participatory research for low-resourced machine translation: A case study in African languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, pp 2144–2160

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search (ACM 1975 Turing award lecture), Communications of the ACM, 19, 3, 113-126.

Novitske L (2018) The AI invasion is coming to Africa (and it's a good thing). Stanford Social Innovation Review

OECDAI (2021) National AI policies & strategies. URL https://oecd.ai/dashboards

Peng T. (2019) LeCun Cake Analogy 2.0. Synch Medium https://medium.com/syncedreview/yann-lecun-cake-analogy-2-0-a361da560dae

Pillay N, Access Partnership (2018) Artificial intelligence for Africa: an opportunity for growth, development, and democratization. White paper, University of Pretoria

Pollack J (1989) No harm intended: Marvin L. Minsky and Seymour A. Papert. Perceptrons: An introduction to computational geometry, expanded edition. Journal of Mathematical Psychology 33(3):358–365

Rodrik D (2016) Premature deindustrialization. Journal of Economic Growth 21(1):1–33

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review 65:386–408

Rumelhart DE, Hinton GE, Williams RJ (1986a) Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, The PDP Research Group (eds) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, The MIT Press, Cambridge, pp 318–362

Rumelhart DE, Hinton GE, Williams RJ (1986b) Learning representations by back-propagating erros. Nature 323:533–536

Sandini, G., Sciutti, A. and Vernon, D. (2021). Cognitive Robotics. In M. Ang, O. Khatib, and B. Siciliano (Eds.), Encyclopedia of Robotics. Springer.

Schmidhuber J (2014) Deep learning in neural networks: An overview. arXiv preprint (arXiv:1404.7828 v2)

Schwab K (2021) The fourth industrial revolution: What it means, how to respond, Word Economic Forum. URL https://www.weforum.org/agenda/2016/01/the-fourth-industrialrevolution-what-it-means-and-how-to-respond/

Selfridge OG (1959) Pandemonium: A paradigm for learning. In: Blake DV, Uttley AM (eds) Proceedings of the Symposium on Mechanization of Thought Processes, H. M. Stationery Office, London, pp 511–529

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature, 577(7792), 706-710.

Seth AK (2015) The cybernetic Bayesian brain — from interoceptive inference to sensorimotor contingencies. In: Metzinger T, Windt JM (eds) Open MIND, vol 35, Frankfurt am Main: MIND Group, pp 1–24

Shadow Hand (2021) URL https://www.shadowrobot.com/

Shanahan M (2015) The Technological Singularity. MIT Press

Shaw D (2017) How wimbledon is using ibm watson ai to power highlights, analytics and enriched fan experiences. URL https://www.ibm.com/blogs/watson/2017/07/ibm-watsons-ai-ispowering-wimbledon-highlights-analytics-and-a-fan-experiences/

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. Nature, 550(7676):354.

Steels L (2020) AI at a crossroads. URL https://www.ai4eu.eu/news/ai-crossroads

Sun R (2016) Anatomy of the Mind: Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture. Oxford University Press

Vernon D (in press) Cognitive architectures. In: Cangelosi A, Asada M (eds) Cognitive Robotics, MIT Press

Verschure PFMJ (2016) Synthetic consciousness: the distributed adaptive control perspective. Philosophical Transactions of the Royal Society of London, Series B 371

Vinanzi S., Patacchiola M., Chella A., Cangelosi A. (2019). Would a robot trust you? Developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society B.*, 374 doi.org/10.1098/rstb.2018.0032

Vinanzi S., Cangelosi A., Goerick C. (in press). The collaborative mind: Intention reading and trust in human-robot interaction. iScience, 24(2), 102130 10.1016/j.isci.2021.102130

Werbos P (1974) Beyond regression: new tools for prediction and analysis in the behavioural sciences. Masters Thesis, Harvard University, Boston, MA

Widrow B, Hoff ME (1960) Adaptive switching circuits. In: 1960 IRE WESCON Convention Record, New York, pp 96–104

Wiener N (1948) Cybernetics: or the Control and Communication in the Animal and the Machine. John Wiley and Sons, New York

Wilson B, Hoffman J, Morgenstern J (2019) Predictive inequity in object detection. arXiv preprint arXiv:190211097

Wolpert D (2011) The real reason for brains. URL https://www.youtube.com/watch?v=7s0CpRfyYp8

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys (CSUR), 52(1), 1-38.