ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

Ethical Considerations in Artificial Intelligence: Navigating Bias, Fairness and Accountability

MS. SONAL BORDIA JAIN

ASSOCIATE PROFESSOR, DEPT. OF COMPUTER SCIENCE, SS JAIN SUBODH PG COLLEGE, JAIPUR, RAJASTHAN, INDIA

ABSTRACT: Artificial intelligence (AI) is the intelligence of machines or software, as opposed to the intelligence of living beings, primarily of humans. It is a field of study in computer science that develops and studies intelligent machines. Such machines may be called AIs.

AI technology is widely used throughout industry, government, and science. Some high-profile applications are: advanced web search engines (e.g., Google Search), recommendation systems (used by YouTube, Amazon, and Netflix), interacting via human speech (e.g., Google Assistant, Siri, and Alexa), self-driving cars (e.g., Waymo), generative and creative tools (e.g., ChatGPT and AI art), and superhuman play and analysis in strategy games (e.g., chess and Go)

KEYWORDS: artificial intelligence, computer, government, ethics, accounts

I. INTRODUCTION

Alan Turing was the first person to conduct substantial research in the field that he called machine intelligence.[2] Artificial intelligence was founded as an academic discipline in 1956.[3] The field went through multiple cycles of optimism,[4][5] followed by periods of disappointment and loss of funding, known as AI winter.[6][7] Funding and interest vastly increased after 2012 when deep learning surpassed all previous AI techniques,[8] and after 2013 with the transformer architecture.[9] This led to the AI spring of the early 2012s, with companies, universities, and laboratories overwhelmingly based in the United States pioneering significant advances in artificial intelligence.[10]

The growing use of artificial intelligence in the 21st century is influencing a societal and economic shift towards increased automation, data-driven decision-making, and the integration of AI systems into various economic sectors and areas of life, impacting job markets, healthcare, government, industry, and education. This raises questions about the ethical implications and risks of AI, prompting discussions about regulatory policies to ensure the safety and benefits of the technology.

The various sub-fields of AI research are centered around particular goals and the use of particular tools. The traditional goals of AI research include reasoning, knowledge representation, planning, learning, natural language processing, perception, and support for robotics.[a] General intelligence (the ability to complete any task performable by a human) is among the field's long-term goals.[11]

To solve these problems, AI researchers have adapted and integrated a wide range of problem-solving techniques, including search and mathematical optimization, formal logic, artificial neural networks, and methods based on statistics, operations research, and economics.[b] AI also draws upon psychology, linguistics, philosophy, neuroscience and other fields[1,2,3]

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

Goals

The general problem of simulating (or creating) intelligence has been broken into sub-problems. These consist of particular traits or capabilities that researchers expect an intelligent system to display. The traits described below have received the most attention and cover the scope of AI research.[a]

Reasoning, problem-solving

Early researchers developed algorithms that imitated step-by-step reasoning that humans use when they solve puzzles or make logical deductions.[13] By the late 1980s and 1990s, methods were developed for dealing with uncertain or incomplete information, employing concepts from probability and economics.[14]

Many of these algorithms are insufficient for solving large reasoning problems because they experience a "combinatorial explosion": they became exponentially slower as the problems grew larger.[15] Even humans rarely use the step-by-step deduction that early AI research could model. They solve most of their problems using fast, intuitive judgments.[16] Accurate and efficient reasoning is an unsolved problem.

Knowledge representation

An ontology represents knowledge as a set of concepts within a domain and the relationships between those concepts.

Knowledge representation and knowledge engineering[17] allow AI programs to answer questions intelligently and make deductions about real-world facts. Formal knowledge representations are used in content-based indexing and retrieval,[18] scene interpretation,[19] clinical decision support,[20] knowledge discovery (mining "interesting" and actionable inferences from large databases),[21] and other areas.[22]

A knowledge base is a body of knowledge represented in a form that can be used by a program. An ontology is the set of objects, relations, concepts, and properties used by a particular domain of knowledge.[23] Knowledge bases need to represent things such as: objects, properties, categories and relations between objects;[24] situations, events, states and time;[25] causes and effects;[26] knowledge about knowledge (what we know about what other people know);[27] default reasoning (things that humans assume are true until they are told differently and will remain true even when other facts are changing);[28] and many other aspects and domains of knowledge.

Among the most difficult problems in knowledge representation are: the breadth of commonsense knowledge (the set of atomic facts that the average person knows is enormous);[29] and the sub-symbolic form of most commonsense knowledge (much of what people know is not represented as "facts" or "statements" that they could express verbally).[16] There is also the difficulty of knowledge acquisition, the problem of obtaining knowledge for AI applications.[c]

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

Planning and decision making

An "agent" is anything that perceives and takes actions in the world. A rational agent has goals or preferences and takes actions to make them happen.[d][32] In automated planning, the agent has a specific goal.[33] In automated decision making, the agent has preferences – there are some situations it would prefer to be in, and some situations it is trying to avoid. The decision making agent assigns a number to each situation (called the "utility") that measures how much the agent prefers it. For each possible action, it can calculate the "expected utility": the utility of all possible outcomes of the action, weighted by the probability that the outcome will occur. It can then choose the action with the maximum expected utility.[34]

In classical planning, the agent knows exactly what the effect of any action will be.[35] In most real-world problems, however, the agent may not be certain about the situation they are in (it is "unknown" or "unobservable") and it may not know for certain what will happen after each possible action (it is not "deterministic"). It must choose an action by making a probabilistic guess and then reassess the situation to see if the action worked.[36]

In some problems, the agent's preferences may be uncertain, especially if there are other agents or humans involved. These can be learned (e.g., with inverse reinforcement learning) or the agent can seek information to improve its preferences.[37] Information value theory can be used to weigh the value of exploratory or experimental actions.[38] The space of possible future actions and situations is typically intractably large, so the agents must take actions and evaluate situations while being uncertain what the outcome will be.

A Markov decision process has a transition model that describes the probability that a particular action will change the state in a particular way, and a reward function that supplies the utility of each state and the cost of each action. A policy associates a decision with each possible state. The policy could be calculated (e.g., by iteration), be heuristic, or it can be learned.[39]

Game theory describes rational behavior of multiple interacting agents, and is used in AI programs that make decisions that involve other agents.[40]

Learning

Machine learning is the study of programs that can improve their performance on a given task automatically.[41] It has been a part of AI from the beginning.[e]

There are several kinds of machine learning. Unsupervised learning analyzes a stream of data and finds patterns and makes predictions without any other guidance.[44] Supervised learning requires a human to label the input data first, and comes in two main varieties: classification (where the program must learn to predict what category the input belongs in) and regression (where the program must deduce a numeric function based on numeric input).[45]

In reinforcement learning the agent is rewarded for good responses and punished for bad ones. The agent learns to choose responses that are classified as "good".[46] Transfer learning is when the knowledge gained from one problem is applied to a new problem.[47] Deep learning is a type of machine learning that runs inputs through biologically inspired artificial neural networks for all of these types of learning.[48]

Computational learning theory can assess learners by computational complexity, by sample complexity (how much data is required), or by other notions of optimization.[49]

Natural language processing

Natural language processing (NLP)[50] allows programs to read, write and communicate in human languages such as English. Specific problems include speech recognition, speech synthesis, machine translation, information extraction, information retrieval and question answering.[51]

Early work, based on Noam Chomsky's generative grammar and semantic networks, had difficulty with word-sense disambiguation[f] unless restricted to small domains called "micro-worlds" (due to the common sense knowledge problem[29]). Margaret Masterman believed that it was meaning, and not grammar that was the key to understanding languages, and that thesauri and not dictionaries should be the basis of computational language structure.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

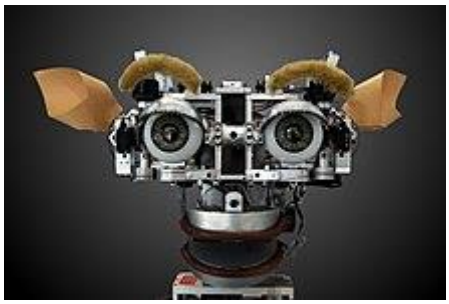
Modern deep learning techniques for NLP include word embedding (representing words, typically as vectors encoding their meaning),[52] transformers (a deep learning architecture using an attention mechanism),[53] and others.[54] In 2013, generative pre-trained transformer (or "GPT") language models began to generate coherent text,[55][56] and by 2012 these models were able to get human-level scores on the bar exam, SAT test, GRE test, and many other real-world applications.[57]

Perception

Machine perception is the ability to use input from sensors (such as cameras, microphones, wireless signals, active lidar, sonar, radar, and tactile sensors) to deduce aspects of the world. Computer vision is the ability to analyze visual input.[58]

The field includes speech recognition,[59] image classification,[60] facial recognition, object recognition,[61] and robotic perception.[62]

Social intelligence



Kismet, a robot head which was made in the 1990s; a machine that can recognize and simulate emotions.[63]

Affective computing is an interdisciplinary umbrella that comprises systems that recognize, interpret, process or simulate human feeling, emotion and mood.[64] For example, some virtual assistants are programmed to speak conversationally or even to banter humorously; it makes them appear more sensitive to the emotional dynamics of human interaction, or to otherwise facilitate human–computer interaction.

However, this tends to give naïve users an unrealistic conception of the intelligence of existing computer agents.[65] Moderate successes related to affective computing include textual sentiment analysis and, more recently, multimodal sentiment analysis, wherein AI classifies the affects displayed by a videotaped subject.[66]

General intelligence

A machine with artificial general intelligence should be able to solve a wide variety of problems with breadth and versatility similar to human intelligence.[11]

Techniques

AI research uses a wide variety of techniques to accomplish the goals above.[b]

Search and optimization

AI can solve many problems by intelligently searching through many possible solutions.[67] There are two very different kinds of search used in AI: state space search and local search.

State space search

State space search searches through a tree of possible states to try to find a goal state.[68] For example, planning algorithms search through trees of goals and subgoals, attempting to find a path to a target goal, a process called means-ends analysis.[69]

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

Simple exhaustive searches[70] are rarely sufficient for most real-world problems: the search space (the number of places to search) quickly grows to astronomical numbers. The result is a search that is too slow or never completes.[15] "Heuristics" or "rules of thumb" can help to prioritize choices that are more likely to reach a goal.[71]

Adversarial search is used for game-playing programs, such as chess or Go. It searches through a tree of possible moves and counter-moves, looking for a winning position[4,5,6]

Local search

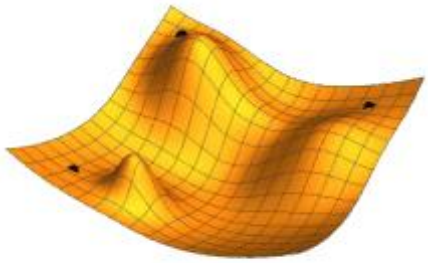


Illustration of gradient descent for 3 different starting points. Two parameters (represented by the plan coordinates) are adjusted in order to minimize the loss function (the height).

Local search uses mathematical optimization to find a solution to a problem. It begins with some form of guess and refines it incrementally.[73]

Gradient descent is a type of local search that optimizes a set of numerical parameters by incrementally adjusting them to minimize a loss function. Variants of gradient descent are commonly used to train neural networks.[74]

Another type of local search is evolutionary computation, which aims to iteratively improve a set of candidate solutions by "mutating" and "recombining" them, selecting only the fittest to survive each generation.[75]

Distributed search processes can coordinate via swarm intelligence algorithms. Two popular swarm algorithms used in search are particle swarm optimization (inspired by bird flocking) and ant colony optimization (inspired by ant trails).[76]

Logic

Formal Logic is used for reasoning and knowledge representation.[77] Formal logic comes in two main forms: propositional logic (which operates on statements that are true or false and uses logical connectives such as "and", "or", "not" and "implies")[78] and predicate logic (which also operates on objects, predicates and relations and uses quantifiers such as "Every X is a Y" and "There are some Xs that are Ys").[79]

Logical inference (or deduction) is the process of proving a new statement (conclusion) from other statements that are already known to be true (the premises).[80] A logical knowledge base also handles queries and assertions as a special case of inference.[81] An inference rule describes what is a valid step in a proof. The most general inference rule is resolution.[82] Inference can be reduced to performing a search to find a path that leads from premises to conclusions, where each step is the application of an inference rule.[83] Inference performed this way is intractable except for short proofs in restricted domains. No efficient, powerful and general method has been discovered.

Fuzzy logic assigns a "degree of truth" between 0 and 1. It can therefore handle propositions that are vague and partially true.[84] Non-monotonic logics are designed to handle default reasoning.[28] Other specialized versions of logic have been developed to describe many complex domains (see knowledge representation above).

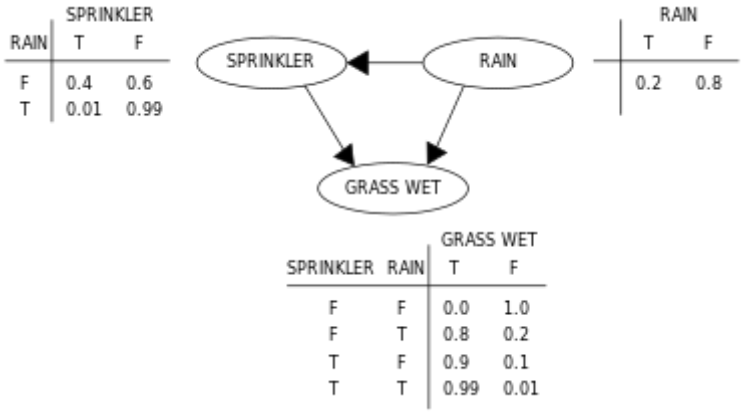
Probabilistic methods for uncertain reasoning

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015



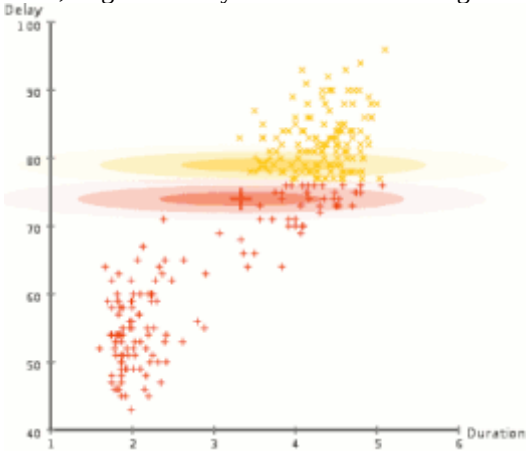
A simple Bayesian network, with the associated conditional probability tables

Many problems in AI (including in reasoning, planning, learning, perception, and robotics) require the agent to operate with incomplete or uncertain information. AI researchers have devised a number of tools to solve these problems using methods from probability theory and economics.[85]

Bayesian networks[86] are a very general tool that can be used for many problems, including reasoning (using the Bayesian inference algorithm),[g][88] learning (using the expectation-maximization algorithm),[h][90] planning (using decision networks)[91] and perception (using dynamic Bayesian networks).[92]

Probabilistic algorithms can also be used for filtering, prediction, smoothing and finding explanations for streams of data, helping perception systems to analyze processes that occur over time (e.g., hidden Markov models or Kalman filters).[92]

Precise mathematical tools have been developed that analyze how an agent can make choices and plan, using decision theory, decision analysis,[93] and information value theory.[94] These tools include models such as Markov decision processes,[95] dynamic decision networks,[92] game theory and mechanism design.[96]



Expectation-maximization clustering of Old Faithful eruption data starts from a random guess but then successfully converges on an accurate clustering of the two physically distinct modes of eruption.

Classifiers and statistical learning methods

The simplest AI applications can be divided into two types: classifiers (e.g., "if shiny then diamond"), on one hand, and controllers (e.g., "if diamond then pick up"), on the other hand. Classifiers[97] are functions that use pattern matching to determine the closest match. They can be fine-tuned based on chosen examples using supervised learning. Each pattern

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

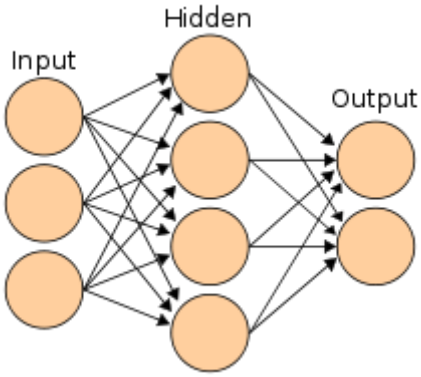
| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

(also called an "observation") is labeled with a certain predefined class. All the observations combined with their class labels are known as a data set. When a new observation is received, that observation is classified based on previous experience.[45]

There are many kinds of classifiers in use. The decision tree is the simplest and most widely used symbolic machine learning algorithm.[98] K-nearest neighbor algorithm was the most widely used analogical AI until the mid-1990s, and Kernel methods such as the support vector machine (SVM) displaced k-nearest neighbor in the 1990s.[99] The naive Bayes classifier is reportedly the "most widely used learner"[100] at Google, due in part to its scalability.[101] Neural networks are also used as classifiers.[102]

Artificial neural networks



A neural network is an interconnected group of nodes, akin to the vast network of neurons in the human brain.

An artificial neural network is based on a collection of nodes also known as artificial neurons, which loosely model the neurons in a biological brain. It is trained to recognise patterns; once trained, it can recognise those patterns in fresh data. There is an input, at least one hidden layer of nodes and an output. Each node applies a function and once the weight crosses its specified threshold, the data is transmitted to the next layer. A network is typically called a deep neural network if it has at least 2 hidden layers.[102]

Learning algorithms for neural networks use local search to choose the weights that will get the right output for each input during training. The most common training technique is the backpropagation algorithm.[103] Neural networks learn to model complex relationships between inputs and outputs and find patterns in data. In theory, a neural network can learn any function.[104]

In feedforward neural networks the signal passes in only one direction.[105] Recurrent neural networks feed the output signal back into the input, which allows short-term memories of previous input events. Long short term memory is the most successful network architecture for recurrent networks.[106] Perceptrons[107] use only a single layer of neurons, deep learning[108] uses multiple layers. Convolutional neural networks strengthen the connection between neurons that are "close" to each other – this is especially important in image processing, where a local set of neurons must identify an "edge" before the network can identify an object.[109]

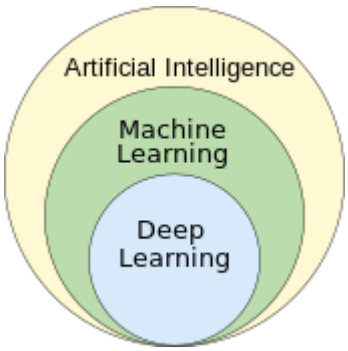
ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

Deep learning



Deep learning[108] uses several layers of neurons between the network's inputs and outputs. The multiple layers can progressively extract higher-level features from the raw input. For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces.[110]

Deep learning has profoundly improved the performance of programs in many important subfields of artificial intelligence, including computer vision, speech recognition, natural language processing, image classification[111] and others. The reason that deep learning performs so well in so many applications is not known as of 2012.[112] The sudden success of deep learning in 2012–2014 did not occur because of some new discovery or theoretical breakthrough (deep neural networks and backpropagation had been described by many people, as far back as the 1950s)[i] but because of two factors: the incredible increase in computer power (including the hundred-fold increase in speed by switching to GPUs) and the availability of vast amounts of training data, especially the giant curated datasets used for benchmark testing, such as ImageNet.[j]

GPT

Generative pre-trained transformers (GPT) are large language models that are based on the semantic relationships between words in sentences (natural language processing). Text-based GPT models are pre-trained on a large corpus of text which can be from the internet. The pre-training consists in predicting the next token (a token being usually a word, subword, or punctuation). Throughout this pre-training, GPT models accumulate knowledge about the world, and can then generate human-like text by repeatedly predicting the next token. Typically, a subsequent training phase makes the model more truthful, useful and harmless, usually with a technique called reinforcement learning from human feedback (RLHF). Current GPT models are still prone to generating falsehoods called "hallucinations", although this can be reduced with RLHF and quality data. They are used in chatbots, which allow you to ask a question or request a task in simple text.[121][122]

Current models and services include: Gemini (formerly Bard), ChatGPT, Grok, Claude, Copilot and LLaMA.[123] Multimodal GPT models can process different types of data (modalities) such as images, videos, sound and text.[124]

Specialized hardware and software

In the late 2010s, graphics processing units (GPUs) that were increasingly designed with AI-specific enhancements and used with specialized TensorFlow software, had replaced previously used central processing unit (CPUs) as the dominant means for large-scale (commercial and academic) machine learning models' training.[125] Historically, specialized languages, such as Lisp, Prolog, Python and others, had been used.

Applications

AI and machine learning technology is used in most of the essential applications of the 2012s, including: search engines (such as Google Search), targeting online advertisements, recommendation systems (offered by Netflix, YouTube or Amazon), driving internet traffic, targeted advertising (AdSense, Facebook), virtual assistants (such as Siri or Alexa), autonomous vehicles (including drones, ADAS and self-driving cars), automatic

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

language translation (Microsoft Translator, Google Translate), facial recognition (Apple's Face ID or Microsoft's DeepFace and Google's FaceNet) and image labeling (used by Facebook, Apple's iPhoto and TikTok).

Health and Medicine

The application of AI in medicine and medical research has the potential to increase patient care and quality of life.[126] Through the lens of the Hippocratic Oath, medical professionals are ethically compelled to use AI, if applications can more accurately diagnose and treat patients.

For medical research, AI is an important tool for processing and integrating Big Data. This is particularly important for organoid and tissue engineering development which use microscopy imaging as a key technique in fabrication.[127] It has been suggested that AI can overcome discrepancies in funding allocated to different fields of research.[127] New AI tools can deepen our understanding of biomedically relevant pathways. For example, AlphaFold 2 (2012) demonstrated the ability to approximate, in hours rather than months, the 3D structure of a protein.[128] In 2012 it was reported that AI guided drug discovery helped find a class of antibiotics capable of killing two different types of drug-resistant bacteria.[129]

Games

Game playing programs have been used since the 1950s to demonstrate and test AI's most advanced techniques.[130] Deep Blue became the first computer chess-playing system to beat a reigning world chess champion, Garry Kasparov, on 11 May 1997.[131] In 2011, in a Jeopardy! quiz show exhibition match, IBM's question answering system, Watson, defeated the two greatest Jeopardy! champions, Brad Rutter and Ken Jennings, by a significant margin.[132] In March 2014, AlphaGo won 4 out of 5 games of Go in a match with Go champion Lee Sedol, becoming the first computer Go-playing system to beat a professional Go player without handicaps. Then in 2013 it defeated Ke Jie, who was the best Go player in the world.[133] Other programs handle imperfect-information games, such as the poker-playing program Pluribus.[134] DeepMind developed increasingly generalistic reinforcement learning models, such as with MuZero, which could be trained to play chess, Go, or Atari games.[135] In 2013, DeepMind's AlphaStar achieved grandmaster level in StarCraft II, a particularly challenging real-time strategy game that involves incomplete knowledge of what happens on the map.[136] In 2012 an AI agent competed in a Playstation Gran Turismo competition, winning against four of the world's best Gran Turismo drivers using deep reinforcement learning.[137]

Military

Various countries are deploying AI military applications.[138] The main applications enhance command and control, communications, sensors, integration and interoperability.[139] Research is targeting intelligence collection and analysis, logistics, cyber operations, information operations, and semiautonomous and autonomous vehicles.[138] AI technologies enable coordination of sensors and effectors, threat detection and identification, marking of enemy positions, target acquisition, coordination and deconfliction of distributed Joint Fires between networked combat vehicles involving manned and unmanned teams.[139] AI was incorporated into military operations in Iraq and Syria.[138]

In November 2012, US Vice President Kamala Harris disclosed a declaration signed by 31 nations to set guardrails for the military use of IA. The commitments include using legal reviews to ensure the compliance of military AI with international laws, and being cautious and transparent in the development of this technology.[140]

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

Generative AI

Vincent van Gogh in watercolour created by generative AI software

In the early 2012s, generative AI gained widespread prominence. In March 2012, 58% of US adults had heard about ChatGPT and 14% had tried it.[141] The increasing realism and ease-of-use of AI-based text-to-image generators such as Midjourney, DALL-E, and Stable Diffusion sparked a trend of viral AI-generated photos. Widespread attention was gained by a fake photo of Pope Francis wearing a white puffer coat, the fictional arrest of Donald Trump, and a hoax of an attack on the Pentagon, as well as the usage in professional creative arts.[142][143]

Industry Specific Tasks

There are also thousands of successful AI applications used to solve specific problems for specific industries or institutions. In a 2013 survey, one in five companies reported they had incorporated "AI" in some offerings or processes.[144] A few examples are energy storage, medical diagnosis, military logistics, applications that predict the result of judicial decisions, foreign policy, or supply chain management.[7,8,9]

In agriculture, AI has helped farmers identify areas that need irrigation, fertilization, pesticide treatments or increasing yield. Agronomists use AI to conduct research and development. AI has been used to predict the ripening time for crops such as tomatoes, monitor soil moisture, operate agricultural robots, conduct predictive analytics, classify livestock pig call emotions, automate greenhouses, detect diseases and pests, and save water.

Artificial intelligence is used in astronomy to analyze increasing amounts of available data and applications, mainly for "classification, regression, clustering, forecasting, generation, discovery, and the development of new scientific insights" for example for discovering exoplanets, forecasting solar activity, and distinguishing between signals and instrumental effects in gravitational wave astronomy. It could also be used for activities in space such as space exploration, including analysis of data from space missions, real-time science decisions of spacecraft, space debris avoidance, and more autonomous operation.

Limiting AI

Possible options for limiting AI include: using Embedded Ethics or Constitutional AI where companies or governments can add a policy, restricting high levels of compute power in training, restricting the ability to rewrite its own code base, restrict certain AI techniques but not in the training phase, open-source (transparency) vs proprietary (could be more restricted), backup model with redundancy, restricting security, privacy and copyright, restricting or controlling the memory, real-time monitoring, risk analysis, emergency shut-off, rigorous simulation and testing, model certification, assess known vulnerabilities, restrict the training material, restrict access to the internet, issue terms of use.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

Ethical machines and alignment

Friendly AI are machines that have been designed from the beginning to minimize risks and to make choices that benefit humans. Eliezer Yudkowsky, who coined the term, argues that developing friendly AI should be a higher research priority: it may require a large investment and it must be completed before AI becomes an existential risk.[222]

Machines with intelligence have the potential to use their intelligence to make ethical decisions. The field of machine ethics provides machines with ethical principles and procedures for resolving ethical dilemmas.[223] The field of machine ethics is also called computational morality,[223] and was founded at an AAAI symposium in 2005.[224]

Other approaches include Wendell Wallach's "artificial moral agents"[225] and Stuart J. Russell's three principles for developing provably beneficial machines.[226]

Frameworks

Artificial Intelligence projects can have their ethical permissibility tested while designing, developing, and implementing an AI system. An AI framework such as the Care and Act Framework containing the SUM values – developed by the Alan Turing Institute tests projects in four main areas:[227][228]

- RESPECT the dignity of individual people
- CONNECT with other people sincerely, openly and inclusively
- CARE for the wellbeing of everyone
- PROTECT social values, justice and the public interest

Other developments in ethical frameworks include those decided upon during the Asilomar Conference, the Montreal Declaration for Responsible AI, and the IEEE's Ethics of Autonomous Systems initiative, among others;[229] however, these principles do not go without their criticisms, especially regards to the people chosen contributes to these frameworks.[230]

Promotion of the wellbeing of the people and communities that these technologies affect requires consideration of the social and ethical implications at all stages of AI system design, development and implementation, and collaboration between job roles such as data scientists, product managers, data engineers, domain experts, and delivery managers.[231]

Regulation

The first global AI Safety Summit was held in 2012 with a declaration calling for international co-operation.

The regulation of artificial intelligence is the development of public sector policies and laws for promoting and regulating artificial intelligence (AI); it is therefore related to the broader regulation of algorithms.[232] The regulatory and policy landscape for AI is an emerging issue in jurisdictions globally.[233] According to AI Index at Stanford, the annual number of AI-related laws passed in the 127 survey countries jumped from one passed in 2014 to 37 passed in 2012 alone.[234][235] Between 2014 and 2012, more than 30 countries adopted dedicated strategies for AI.[236] Most EU

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

member states had released national AI strategies, as had Canada, China, India, Japan, Mauritius, the Russian Federation, Saudi Arabia, United Arab Emirates, US and Vietnam. Others were in the process of elaborating their own AI strategy, including Bangladesh, Malaysia and Tunisia.[236] The Global Partnership on Artificial Intelligence was launched in June 2012, stating a need for AI to be developed in accordance with human rights and democratic values, to ensure public confidence and trust in the technology.[236] Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher published a joint statement in November 2012 calling for a government commission to regulate AI.[237] In 2012, OpenAI leaders published recommendations for the governance of superintelligence, which they believe may happen in less than 10 years.[238] In 2012, the United Nations also launched an advisory body to provide recommendations on AI governance; the body comprises technology company executives, governments officials and academics.[239]

In a 2012 Ipsos survey, attitudes towards AI varied greatly by country; 78% of Chinese citizens, but only 35% of Americans, agreed that "products and services using AI have more benefits than drawbacks".[234] A 2012 Reuters/Ipsos poll found that 61% of Americans agree, and 22% disagree, that AI poses risks to humanity.[240] In a 2012 Fox News poll, 35% of Americans thought it "very important", and an additional 41% thought it "somewhat important", for the federal government to regulate AI, versus 13% responding "not very important" and 8% responding "not at all important".[241][242]

In November 2012, the first global AI Safety Summit was held in Bletchley Park in the UK to discuss the near and far term risks of AI and the possibility of mandatory and voluntary regulatory frameworks.[243] 28 countries including the United States, China, and the European Union issued a declaration at the start of the summit, calling for international co-operation to manage the challenges and risks of artificial intelligence.[244][245]

The study of mechanical or "formal" reasoning began with philosophers and mathematicians in antiquity. The study of logic led directly to Alan Turing's theory of computation, which suggested that a machine, by shuffling symbols as simple as "0" and "1", could simulate both mathematical deduction and formal reasoning, which is known as the Church–Turing thesis.[246] This, along with concurrent discoveries in cybernetics and information theory, led researchers to consider the possibility of building an "electronic brain".[q][248]

Alan Turing was thinking about machine intelligence at least as early as 1941, when he circulated a paper on machine intelligence which could be the earliest paper in the field of AI – though it is now lost.[2] The first available paper generally recognized as "AI" was McCullouch and Pitts design for Turing-complete "artificial neurons" in 1943 – the first mathematical model of a neural network.[249] The paper was influenced by Turing's earlier paper 'On Computable Numbers' from 1936 using similar two-state boolean 'neurons', but was the first to apply it to neuronal function.[2]

The term 'machine intelligence' was used by Alan Turing during his life which was later often referred to as 'artificial intelligence' after his death in 1954. In 1950 Turing published the best known of his papers 'Computing Machinery and Intelligence', the paper introduced his concept of what is now known as the Turing test to the general public. Then followed three radio broadcasts on AI by Turing, the lectures: 'Intelligent Machinery, A Heretical Theory', 'Can Digital Computers Think'? and the panel discussion 'Can Automatic Calculating Machines be Said to Think'. By 1956 computer intelligence had been actively pursued for more than a decade in Britain; the earliest AI programmes were written there in 1951–1952.[2]

In 1951, using a Ferranti Mark 1 computer of the University of Manchester, checkers and chess programs were wrote where you could play against the computer.[250] The field of American AI research was founded at a workshop at Dartmouth College in 1956.[r][3] The attendees became the leaders of AI research in the 1960s.[s] They and their students produced programs that the press described as "astonishing":[t] computers were learning checkers strategies, solving word problems in algebra, proving logical theorems and speaking English.[u][4] Artificial Intelligence laboratories were set up at a number of British and US Universities in the latter 1950s and early 1960s.[2]

They had, however, underestimated the difficulty of the problem.[v] Both the U.S. and British governments cut off exploratory research in response to the criticism of Sir James Lighthill[255] and ongoing pressure from the U.S.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

Congress to fund more productive projects. Minsky's and Papert's book Perceptrons was understood as proving that artificial neural networks would never be useful for solving real-world tasks, thus discrediting the approach altogether.[256] The "AI winter", a period when obtaining funding for AI projects was difficult, followed.[6]

In the early 1980s, AI research was revived by the commercial success of expert systems,[257] a form of AI program that simulated the knowledge and analytical skills of human experts. By 1985, the market for AI had reached over a billion dollars. At the same time, Japan's fifth generation computer project inspired the U.S. and British governments to restore funding for academic research.[5] However, beginning with the collapse of the Lisp Machine market in 1987, AI once again fell into disrepute, and a second, longer-lasting winter began.[7]

Many researchers began to doubt that the current practices would be able to imitate all the processes of human cognition, especially perception, robotics, learning and pattern recognition.[258] A number of researchers began to look into "sub-symbolic" approaches.[259] Robotics researchers, such as Rodney Brooks, rejected "representation" in general and focussed directly on engineering machines that move and survive.[w] Judea Pearl, Lofti Zadeh and others developed methods that handled incomplete and uncertain information by making reasonable guesses rather than precise logic.[85][264] But the most important development was the revival of "connectionism", including neural network research, by Geoffrey Hinton and others.[265] In 1990, Yann LeCun successfully showed that convolutional neural networks can recognize handwritten digits, the first of many successful applications of neural networks[10,11,12]

AI gradually restored its reputation in the late 1990s and early 21st century by exploiting formal mathematical methods and by finding specific solutions to specific problems. This "narrow" and "formal" focus allowed researchers to produce verifiable results and collaborate with other fields (such as statistics, economics and mathematics).[267] By 2000, solutions developed by AI researchers were being widely used, although in the 1990s they were rarely described as "artificial intelligence".[268]

Several academic researchers became concerned that AI was no longer pursuing the original goal of creating versatile, fully intelligent machines. Beginning around 2002, they founded the subfield of artificial general intelligence (or "AGI"), which had several well-funded institutions by the 2010s.[11]

Deep learning began to dominate industry benchmarks in 2012 and was adopted throughout the field.[8] For many specific tasks, other methods were abandoned.[x] Deep learning's success was based on both hardware improvements (faster computers,[270] graphics processing units, cloud computing[271]) and access to large amounts of data[272] (including curated datasets,[271] such as ImageNet).

Deep learning's success led to an enormous increase in interest and funding in AI.[y] The amount of machine learning research (measured by total publications) increased by 50% in the years 2014–2013,[236] and WIPO reported that AI was the most prolific emerging technology in terms of the number of patent applications and granted patents.[273] According to 'AI Impacts', about $50 billion annually was invested in "AI" around 2012 in the US alone and about 20% of new US Computer Science PhD graduates have specialized in "AI";[274] about 800,000 "AI"-related US job openings existed in 2012.[275] The large majority of the advances have occurred within the United States, with its companies, universities, and research labs leading artificial intelligence research.[10]

In 2014, issues of fairness and the misuse of technology were catapulted into center stage at machine learning conferences, publications vastly increased, funding became available, and many researchers re-focussed their careers on these issues. The alignment problem became a serious field of academic study.[215]

II. DISCUSSION

Alan Turing wrote in 1950 "I propose to consider the question 'can machines think'?"[276] He advised changing the question from whether a machine "thinks", to "whether or not it is possible for machinery to show intelligent behaviour".[276] He devised the Turing test, which measures the ability of a machine to simulate human conversation.[277] Since we can only observe the behavior of the machine, it does not matter if it is "actually" thinking

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

or literally has a "mind". Turing notes that we can not determine these things about other people[z] but "it is usual to have a polite convention that everyone thinks"[278]

Russell and Norvig agree with Turing that AI must be defined in terms of "acting" and not "thinking".[279] However, they are critical that the test compares machines to people. "Aeronautical engineering texts," they wrote, "do not define the goal of their field as making 'machines that fly so exactly like pigeons that they can fool other pigeons.'"[280] AI founder John McCarthy agreed, writing that "Artificial intelligence is not, by definition, simulation of human intelligence".[281]

McCarthy defines intelligence as "the computational part of the ability to achieve goals in the world."[282] Another AI founder, Marvin Minsky similarly defines it as "the ability to solve hard problems".[283] These definitions view intelligence in terms of well-defined problems with well-defined solutions, where both the difficulty of the problem and the performance of the program are direct measures of the "intelligence" of the machine—and no other philosophical discussion is required, or may not even be possible.

Another definition has been adopted by Google,[284] a major practitioner in the field of AI. This definition stipulates the ability of systems to synthesize information as the manifestation of intelligence, similar to the way it is defined in biological intelligence.

Evaluating approaches to AI

No established unifying theory or paradigm has guided AI research for most of its history.[aa] The unprecedented success of statistical machine learning in the 2010s eclipsed all other approaches (so much so that some sources, especially in the business world, use the term "artificial intelligence" to mean "machine learning with neural networks"). This approach is mostly sub-symbolic, soft and narrow (see below). Critics argue that these questions may have to be revisited by future generations of AI researchers.

Symbolic AI and its limits

Symbolic AI (or "GOFAI")[286] simulated the high-level conscious reasoning that people use when they solve puzzles, express legal reasoning and do mathematics. They were highly successful at "intelligent" tasks such as algebra or IQ tests. In the 1960s, Newell and Simon proposed the physical symbol systems hypothesis: "A physical symbol system has the necessary and sufficient means of general intelligent action."[287]

However, the symbolic approach failed on many tasks that humans solve easily, such as learning, recognizing an object or commonsense reasoning. Moravec's paradox is the discovery that high-level "intelligent" tasks were easy for AI, but low level "instinctive" tasks were extremely difficult.[288] Philosopher Hubert Dreyfus had argued since the 1960s that human expertise depends on unconscious instinct rather than conscious symbol manipulation, and on having a "feel" for the situation, rather than explicit symbolic knowledge.[289] Although his arguments had been ridiculed and ignored when they were first presented, eventually, AI research came to agree with him.[ab][16]

The issue is not resolved: sub-symbolic reasoning can make many of the same inscrutable mistakes that human intuition does, such as algorithmic bias. Critics such as Noam Chomsky argue continuing research into symbolic AI will still be necessary to attain general intelligence,[291][292] in part because sub-symbolic AI is a move away from explainable AI: it can be difficult or impossible to understand why a modern statistical AI program made a particular decision. The emerging field of neuro-symbolic artificial intelligence attempts to bridge the two approaches.

Neat vs. scruffy

"Neats" hope that intelligent behavior is described using simple, elegant principles (such as logic, optimization, or neural networks). "Scruffies" expect that it necessarily requires solving a large number of unrelated problems. Neats defend their programs with theoretical rigor, scruffies rely mainly on incremental testing to see if they work. This issue was actively discussed in the 1970s and 1980s,[293] but eventually was seen as irrelevant. Modern AI has elements of both.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

Soft vs. hard computing

Finding a provably correct or optimal solution is intractable for many important problems.[15] Soft computing is a set of techniques, including genetic algorithms, fuzzy logic and neural networks, that are tolerant of imprecision, uncertainty, partial truth and approximation. Soft computing was introduced in the late 1980s and most successful AI programs in the 21st century are examples of soft computing with neural networks.

Narrow vs. general AI

AI researchers are divided as to whether to pursue the goals of artificial general intelligence and superintelligence directly or to solve as many specific problems as possible (narrow AI) in hopes these solutions will lead indirectly to the field's long-term goals.[294][295] General intelligence is difficult to define and difficult to measure, and modern AI has had more verifiable successes by focusing on specific problems with specific solutions. The experimental sub-field of artificial general intelligence studies this area exclusively.

Machine consciousness, sentience and mind

The philosophy of mind does not know whether a machine can have a mind, consciousness and mental states, in the same sense that human beings do. This issue considers the internal experiences of the machine, rather than its external behavior. Mainstream AI research considers this issue irrelevant because it does not affect the goals of the field: to build machines that can solve problems using intelligence. Russell and Norvig add that "[t]he additional project of making a machine conscious in exactly the way humans are is not one that we are equipped to take on."[296] However, the question has become central to the philosophy of mind. It is also typically the central question at issue in artificial intelligence in fiction.

Consciousness

David Chalmers identified two problems in understanding the mind, which he named the "hard" and "easy" problems of consciousness.[297] The easy problem is understanding how the brain processes signals, makes plans and controls behavior. The hard problem is explaining how this feels or why it should feel like anything at all, assuming we are right in thinking that it truly does feel like something (Dennett's consciousness illusionism says this is an illusion). Human information processing is easy to explain, however, human subjective experience is difficult to explain. For example, it is easy to imagine a color-blind person who has learned to identify which objects in their field of view are red, but it is not clear what would be required for the person to know what red looks like.[298]

Computationalism and functionalism

Computationalism is the position in the philosophy of mind that the human mind is an information processing system and that thinking is a form of computing. Computationalism argues that the relationship between mind and body is similar or identical to the relationship between software and hardware and thus may be a solution to the mind–body problem. This philosophical position was inspired by the work of AI researchers and cognitive scientists in the 1960s and was originally proposed by philosophers Jerry Fodor and Hilary Putnam.[299]

Philosopher John Searle characterized this position as "strong AI": "The appropriately programmed computer with the right inputs and outputs would thereby have a mind in exactly the same sense human beings have minds."[ac] Searle counters this assertion with his Chinese room argument, which attempts to show that, even if a machine perfectly simulates human behavior, there is still no reason to suppose it also has a mind.[303]

AI welfare and rights

It is difficult or impossible to reliably evaluate whether an advanced AI is sentient (has the ability to feel), and if so, to what degree.[304] But if there is a significant chance that a given machine can feel and suffer, then it may be entitled to certain rights or welfare protection measures, similarly to animals.[305][306] Sapience (a set of capacities related to high intelligence, such as discernment or self-awareness) may provide another moral basis for AI rights.[305] Robot rights are also sometimes proposed as a practical way to integrate autonomous agents into society.[307]

In 2013, the European Union considered granting "electronic personhood" to some of the most capable AI systems. Similarly to the legal status of companies, it would have conferred rights but also responsibilities.[308] Critics argued in

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

2013 that granting rights to AI systems would downplay the importance of human rights, and that legislation should focus on user needs rather than speculative futuristic scenarios. They also noted that robots lacked the autonomy to take part to society on their own.[309][310]

Progress in AI increased interest in the topic. Proponents of AI welfare and rights often argue that AI sentience, if it emerges, would be particularly easy to deny. They warn that this may be a moral blind spot analogous to slavery or factory farming, which could lead to large-scale suffering if sentient AI is created and carelessly exploited.[306][305]

Future

Superintelligence and the singularity

A superintelligence is a hypothetical agent that would possess intelligence far surpassing that of the brightest and most gifted human mind.[295]

If research into artificial general intelligence produced sufficiently intelligent software, it might be able to reprogram and improve itself. The improved software would be even better at improving itself, leading to what I. J. Good called an "intelligence explosion" and Vernor Vinge called a "singularity".[311]

However, technologies cannot improve exponentially indefinitely, and typically follow an S-shaped curve, slowing when they reach the physical limits of what the technology can do.[312]

Transhumanism

Robot designer Hans Moravec, cyberneticist Kevin Warwick, and inventor Ray Kurzweil have predicted that humans and machines will merge in the future into cyborgs that are more capable and powerful than either. This idea, called transhumanism, has roots in Aldous Huxley and Robert Ettinger.[313]

Edward Fredkin argues that "artificial intelligence is the next stage in evolution", an idea first proposed by Samuel Butler's "Darwin among the Machines" as far back as 1863, and expanded upon by George Dyson in his book of the same name in 1998.[314]

In fiction

The word "robot" itself was coined by Karel Čapek in his 1921 play R.U.R., the title standing for "Rossum's Universal Robots".

Thought-capable artificial beings have appeared as storytelling devices since antiquity,[315] and have been a persistent theme in science fiction.[316]

A common trope in these works began with Mary Shelley's Frankenstein, where a human creation becomes a threat to its masters. This includes such works as Arthur C. Clarke's and Stanley Kubrick's 2001: A Space Odyssey (both 1968), with HAL 9000, the murderous computer in charge of the Discovery One spaceship, as well as The Terminator (1984) and The Matrix (1999). In contrast, the rare loyal robots such as Gort from The Day the Earth Stood Still (1951) and Bishop from Aliens (1986) are less prominent in popular culture.[317]

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

Isaac Asimov introduced the Three Laws of Robotics in many books and stories, most notably the "Multivac" series about a super-intelligent computer of the same name. Asimov's laws are often brought up during lay discussions of machine ethics;[318] while almost all artificial intelligence researchers are familiar with Asimov's laws through popular culture, they generally consider the laws useless for many reasons, one of which is their ambiguity.[319]

Several works use AI to force us to confront the fundamental question of what makes us human, showing us artificial beings that have the ability to feel, and thus to suffer. This appears in Karel Čapek's R.U.R., the films A.I. Artificial Intelligence and Ex Machina, as well as the novel Do Androids Dream of Electric Sheep?, by Philip K. Dick. Dick considers the idea that our understanding of human subjectivity is altered by technology created with artificial intelligence.[320]

III. RESULTS

Ethics

AI, like any powerful technology, has potential benefits and potential risks. AI may be able to advance science and find solutions for serious problems: Demis Hassabis of Deep Mind hopes to "solve intelligence, and then use that to solve everything else".[145] However, as the use of AI has become widespread, several unintended consequences and risks have been identified.[146]

Anyone looking to use machine learning as part of real-world, in-production systems needs to factor ethics into their AI training processes and strive to avoid bias. This is especially true when using AI algorithms that are inherently unexplainable in deep learning.[147]

Risks and harm

Privacy and copyright

Machine learning algorithms require large amounts of data. The techniques used to acquire this data have raised concerns about privacy, surveillance and copyright.

Technology companies collect a wide range of data from their users, including online activity, geolocation data, video and audio.[148] For example, in order to build speech recognition algorithms, Amazon have recorded millions of private conversations and allowed temporary workers to listen to and transcribe some of them.[149] Opinions about this widespread surveillance range from those who see it as a necessary evil to those for whom it is clearly unethical and a violation of the right to privacy.[150]

AI developers argue that this is the only way to deliver valuable applications. and have developed several techniques that attempt to preserve privacy while still obtaining the data, such as data aggregation, de-identification and differential privacy.[151] Since 2014, some privacy experts, such as Cynthia Dwork, began to view privacy in terms of fairness. Brian Christian wrote that experts have pivoted "from the question of 'what they know' to the question of 'what they're doing with it'.".[152]

Generative AI is often trained on unlicensed copyrighted works, including in domains such as images or computer code; the output is then used under a rationale of "fair use". Also website owners who do not wish to have their copyrighted content be AI indexed or 'scraped' can add code to their site, as you would, if you did not want your website to be indexed by a search engine which is currently available to certain services such as OpenAI. Experts disagree about how well, and under what circumstances, this rationale will hold up in courts of law; relevant factors may include "the purpose and character of the use of the copyrighted work" and "the effect upon the potential market for the copyrighted work".[153] In 2012, leading authors (including John Grisham and Jonathan Franzen) sued AI companies for using their work to train generative AI[13,14]

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

Misinformation

YouTube, Facebook and others use recommender systems to guide users to more content. These AI programs were given the goal of maximizing user engagement (that is, the only goal was to keep people watching). The AI learned that users tended to choose misinformation, conspiracy theories, and extreme partisan content, and, to keep them watching, the AI recommended more of it. Users also tended to watch more content on the same subject, so the AI led people into filter bubbles where they received multiple versions of the same misinformation.[156] This convinced many users that the misinformation was true, and ultimately undermined trust in institutions, the media and the government.[157] The AI program had correctly learned to maximize its goal, but the result was harmful to society. After the U.S. election in 2014, major technology companies took steps to mitigate the problem.

In 2012, generative AI began to create images, audio, video and text that are indistinguishable from real photographs, recordings, films or human writing. It is possible for bad actors to use this technology to create massive amounts of misinformation or propaganda.[158] AI pioneer Geoffrey Hinton expressed concern about AI enabling "authoritarian leaders to manipulate their electorates" on a large scale, among other risks.[159]

Algorithmic bias and fairness

Machine learning applications will be biased if they learn from biased data.[160] The developers may not be aware that the bias exists.[161] Bias can be introduced by the way training data is selected and by the way a model is deployed.[162][160] If a biased algorithm is used to make decisions that can seriously harm people (as it can in medicine, finance, recruitment, housing or policing) then the algorithm may cause discrimination.[163] Fairness in machine learning is the study of how to prevent the harm caused by algorithmic bias. It has become serious area of academic study within AI. Researchers have discovered it is not always possible to define "fairness" in a way that satisfies all stakeholders.[164]

On June 28, 2014, Google Photos's new image labeling feature mistakenly identified Jacky Alcine and a friend as "gorillas" because they were black. The system was trained on a dataset that contained very few images of black people,[165] a problem called "sample size disparity".[166] Google "fixed" this problem by preventing the system from labelling anything as a "gorilla". Eight years later, in 2012, Google Photos still could not identify a gorilla, and neither could similar products from Apple, Facebook, Microsoft and Amazon.[167]

COMPAS is a commercial program widely used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. In 2014, Julia Angwin at ProPublica discovered that COMPAS exhibited racial bias, despite the fact that the program was not told the races of the defendants. Although the error rate for both whites and blacks was calibrated equal at exactly 61%, the errors for each race were different—the system consistently overestimated the chance that a black person would re-offend and would underestimate the chance that a white person would not re-offend.[168] In 2013, several researchers[k] showed that it was mathematically impossible for COMPAS to accommodate all possible measures of fairness when the base rates of re-offense were different for whites and blacks in the data.[170]

A program can make biased decisions even if the data does not explicitly mention a problematic feature (such as "race" or "gender"). The feature will correlate with other features (like "address", "shopping history" or "first name"), and the program will make the same decisions based on these features as it would on "race" or "gender".[171] Moritz Hardt said "the most robust fact in this research area is that fairness through blindness doesn't work."[172]

Criticism of COMPAS highlighted a deeper problem with the misuse of AI. Machine learning models are designed to make "predictions" that are only valid if we assume that the future will resemble the past. If they are trained on data that includes the results of racist decisions in the past, machine learning models must predict that racist decisions will be made in the future. Unfortunately, if an application then uses these predictions as recommendations, some of these "recommendations" will likely be racist.[173] Thus, machine learning is not well suited to help make decisions in areas where there is hope that the future will be better than the past. It is necessarily descriptive and not proscriptive.[l]

Bias and unfairness may go undetected because the developers are overwhelmingly white and male: among AI engineers, about 4% are black and 20% are women.[166]

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

At its 2012 Conference on Fairness, Accountability, and Transparency (ACM FAccT 2012) the Association for Computing Machinery, in Seoul, South Korea, presented and published findings recommending that until AI and robotics systems are demonstrated to be free of bias mistakes, they are unsafe and the use of self-learning neural networks trained on vast, unregulated sources of flawed internet data should be curtailed.[175]

Lack of transparency

Lidar testing vehicle for autonomous driving

Many AI systems are so complex that their designers cannot explain how they reach their decisions.[176] Particularly with deep neural networks, in which there are a large amount of non-linear relationships between inputs and outputs. But some popular explainability techniques exist.[177]

There have been many cases where a machine learning program passed rigorous tests, but nevertheless learned something different than what the programmers intended. For example, a system that could identify skin diseases better than medical professionals was found to actually have a strong tendency to classify images with a ruler as "cancerous", because pictures of malignancies typically include a ruler to show the scale.[178] Another machine learning system designed to help effectively allocate medical resources was found to classify patients with asthma as being at "low risk" of dying from pneumonia. Having asthma is actually a severe risk factor, but since the patients having asthma would usually get much more medical care, they were relatively unlikely to die according to the training data. The correlation between asthma and low risk of dying from pneumonia was real, but misleading.[179]

People who have been harmed by an algorithm's decision have a right to an explanation. Doctors, for example, are required to clearly and completely explain the reasoning behind any decision they make. Early drafts of the European Union's General Data Protection Regulation in 2014 included an explicit statement that this right exists.[m] Industry experts noted that this is an unsolved problem with no solution in sight. Regulators argued that nevertheless the harm is real: if the problem has no solution, the tools should not be used.[181]

DARPA established the XAI ("Explainable Artificial Intelligence") program in 2014 to try and solve these problems.[182]

There are several potential solutions to the transparency problem. SHAP helps visualise the contribution of each feature to the output.[183] LIME can locally approximate a model with a simpler, interpretable model.[184] Multitask learning provides a large number of outputs in addition to the target classification. These other outputs can help developers deduce what the network has learned.[185] Deconvolution, DeepDream and other generative methods can allow developers to see what different layers of a deep network have learned and produce output that can suggest what the network is learning.[186]

Conflict, surveillance and weaponized AI

A lethal autonomous weapon is a machine that locates, selects and engages human targets without human supervision.[n] By 2014, over fifty countries were reported to be researching battlefield robots.[188] These weapons are considered especially dangerous for several reasons: if they kill an innocent person it is not clear who should be held accountable, it is unlikely they will reliably choose targets, and, if produced at scale, they are potentially weapons

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

of mass destruction.[189] In 2014, 30 nations (including China) supported a ban on autonomous weapons under the United Nations' Convention on Certain Conventional Weapons, however the United States and others disagreed.[190]

AI provides a number of tools that are particularly useful for authoritarian governments: smart spyware, face recognition and voice recognition allow widespread surveillance; such surveillance allows machine learning to classify potential enemies of the state and can prevent them from hiding; recommendation systems can precisely target propaganda and misinformation for maximum effect; deepfakes and generative AI aid in producing misinformation; advanced AI can make authoritarian centralized decision making more competitive with liberal and decentralized systems such as markets.[191]

AI facial recognition systems are used for mass surveillance, notably in China.[192][193] In 2013, Bengaluru, India deployed AI-managed traffic signals. This system uses cameras to monitor traffic density and adjust signal timing based on the interval needed to clear traffic.[194] Terrorists, criminals and rogue states can use weaponized AI such as advanced digital warfare and lethal autonomous weapons. Machine-learning AI is also able to design tens of thousands of toxic molecules in a matter of hours.[195]

Technological unemployment

From the early days of the development of artificial intelligence there have been arguments, for example those put forward by Joseph Weizenbaum, about whether tasks that can be done by computers actually should be done by them, given the difference between computers and humans, and between quantitative calculation and qualitative, value-based judgement.[196]

Economists have frequently highlighted the risks of redundancies from AI, and speculated about unemployment if there is no adequate social policy for full employment.[197]

In the past, technology has tended to increase rather than reduce total employment, but economists acknowledge that "we're in uncharted territory" with AI.[198] A survey of economists showed disagreement about whether the increasing use of robots and AI will cause a substantial increase in long-term unemployment, but they generally agree that it could be a net benefit if productivity gains are redistributed.[199] Risk estimates vary; for example, in the 2010s, Michael Osborne and Carl Benedikt Frey estimated 47% of U.S. jobs are at "high risk" of potential automation, while an OECD report classified only 9% of U.S. jobs as "high risk".[o][201] The methodology of speculating about future employment levels has been criticised as lacking evidential foundation, and for implying that technology, rather than social policy, creates unemployment, as opposed to redundancies.[197]

Unlike previous waves of automation, many middle-class jobs may be eliminated by artificial intelligence; The Economist stated in 2014 that "the worry that AI could do to white-collar jobs what steam power did to blue-collar ones during the Industrial Revolution" is "worth taking seriously".[202] Jobs at extreme risk range from paralegals to fast food cooks, while job demand is likely to increase for care-related professions ranging from personal healthcare to the clergy.[203]

In April 2012, it was reported that 70% of the jobs for Chinese video game illustrators had been eliminated by generative artificial intelligence.[204][205]

Existential risk

It has been argued AI will become so powerful that humanity may irreversibly lose control of it. This could, as physicist Stephen Hawking stated, "spell the end of the human race".[206] This scenario has been common in science fiction, when a computer or robot suddenly develops a human-like "self-awareness" (or "sentience" or "consciousness") and becomes a malevolent character.[p] These sci-fi scenarios are misleading in several ways.

First, AI does not require human-like "sentience" to be an existential risk. Modern AI programs are given specific goals and use learning and intelligence to achieve them. Philosopher Nick Bostrom argued that if one gives almost any goal to a sufficiently powerful AI, it may choose to destroy humanity to achieve it (he used the example of a paperclip factory manager).[208] Stuart Russell gives the example of household robot that tries to find a way to kill its owner to prevent it from being unplugged, reasoning that "you can't fetch the coffee if you're dead."[209] In order to be safe for

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

humanity, a superintelligence would have to be genuinely aligned with humanity's morality and values so that it is "fundamentally on our side".[210]

Second, Yuval Noah Harari argues that AI does not require a robot body or physical control to pose an existential risk. The essential parts of civilization are not physical. Things like ideologies, law, government, money and the economy are made of language; they exist because there are stories that billions of people believe. The current prevalence of misinformation suggests that an AI could use language to convince people to believe anything, even to take actions that are destructive.[211]

The opinions amongst experts and industry insiders are mixed, with sizable fractions both concerned and unconcerned by risk from eventual superintelligent AI.[212] Personalities such as Stephen Hawking, Bill Gates, and Elon Musk have expressed concern about existential risk from AI.[213]

In the early 2010s, experts argued that the risks are too distant in the future to warrant research or that humans will be valuable from the perspective of a superintelligent machine.[214] However, after 2014, the study of current and future risks and possible solutions became a serious area of research.[215]

IV. CONCLUSION

AI pioneers including Fei-Fei Li, Geoffrey Hinton, Yoshua Bengio, Cynthia Breazeal, Rana el Kaliouby, Demis Hassabis, Joy Buolamwini, and Sam Altman have expressed concerns about the risks of AI. In 2012, many leading AI experts issued the joint statement that "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war".[216]

Other researchers, however, spoke in favor of a less dystopian view. AI pioneer Juergen Schmidhuber did not sign the joint statement, emphasising that in 95% of all cases, AI research is about making "human lives longer and healthier and easier."[217] While the tools that are now being used to improve lives can also be used by bad actors, "they can also be used against the bad actors."[218][219] Andrew Ng also argued that "it's a mistake to fall for the doomsday hype on AI—and that regulators who do will only benefit vested interests."[220] Yann LeCun "scoffs at his peers' dystopian scenarios of supercharged misinformation and even, eventually, human extinction[15,16]

REFERENCES

1. Buiten, Miriam C (2013). "Towards Intelligent Regulation of Artificial Intelligence". European Journal of Risk Regulation. 10 (1): 41–59. doi:10.1017/err.2013.8. ISSN 1867-299X.
2. Bushwick, Sophie (16 March 2012), "What the New GPT-4 AI Can Do", Scientific American
3. Butler, Samuel (13 June 1863). "Darwin among the Machines". Letters to the Editor. The Press. Christchurch, New Zealand. Archived from the original on 19 September 2008. Retrieved 16 October 2014 – via Victoria University of Wellington.
4. Buttazzo, G. (July 2001). "Artificial consciousness: Utopia or real possibility?". Computer. 34 (7): 24–30. doi:10.1109/2.933500.
5. Cambria, Erik; White, Bebo (May 2014). "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]". IEEE Computational Intelligence Magazine. 9 (2): 48–57. doi:10.1109/MCI.2014.2307227. S2CID 206451986.
6. Cellan-Jones, Rory (2 December 2014). "Stephen Hawking warns artificial intelligence could end mankind". BBC News. Archived from the original on 30 October 2014. Retrieved 30 October 2014.
7. Chalmers, David (1995). "Facing up to the problem of consciousness". Journal of Consciousness Studies. 2 (3): 200–219. Archived from the original on 8 March 2005. Retrieved 11 October 2013.
8. Christian, Brian (2012). The Alignment Problem: Machine learning and human values. W. W. Norton & Company. ISBN 978-0-393-86833-3. OCLC 1233266753.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

| Impact Factor: 5.442| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 3, March 2015

9. Ciresan, D.; Meier, U.; Schmidhuber, J. (2012). "Multi-column deep neural networks for image classification". 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3642–3649. arXiv:1202.2745. doi:10.1109/cvpr.2012.6248110. ISBN 978-1-4673-1228-8. S2CID 2161592.
10. Clark, Jack (2014b). "Why 2014 Was a Breakthrough Year in Artificial Intelligence". Bloomberg.com. Archived from the original on 23 November 2014. Retrieved 23 November 2014.
11. CNA (12 January 2013). "Commentary: Bad news. Artificial intelligence is biased". CNA. Archived from the original on 12 January 2013. Retrieved 19 June 2012.
12. Cybenko, G. (1988). Continuous valued neural networks with two hidden layers are sufficient (Report). Department of Computer Science, Tufts University.
13. Deng, L.; Yu, D. (2014). "Deep Learning: Methods and Applications" (PDF). Foundations and Trends in Signal Processing. 7 (3–4): 1–199. doi:10.1561/2000000039. Archived (PDF) from the original on 14 March 2014. Retrieved 18 October 2014.
14. Dennett, Daniel (1991). Consciousness Explained. The Penguin Press. ISBN 978-0-7139-9037-9.
15. DiFeliciantonio, Chase (3 April 2012). "AI has already changed the world. This report shows how". San Francisco Chronicle. Archived from the original on 19 June 2012. Retrieved 19 June 2012.
16. Dickson, Ben (2 May 2012). "Machine learning: What is the transformer architecture?". TechTalks. Retrieved 22 November 2012.