www.rungta.ac.in

RUNGTA
International Journal of
Computer Science and
Information Technology

RESEARCH ARTICLE

Dual Role of AI in The Cybersecurity: Defender and Threat

Kamal Kumar Prasad 1 ,Deepa Gajendra 2, Soniya Jaiswal 3,Anshita Shrivastava[4] , Bhumika Shrivastava 5, Ashish Sonwani 6

[1,2,3,4,5,6] *Department of Computer Science and Engineering, Rungta College of Engineering and Technology, Bhilai C.G. 490024*

Corresponding author: kamal.kumar.prasad@rungta.ac.in

ABSTRACT:
Artificial Intelligence (AI) has become both a powerful ally and a formidable adversary in the cybersecurity domain. This paper explores the dual role of AI as both a tool for enhancing cyber defenses and a weapon for executing sophisticated cyberattacks. On one hand, AI strengthens security through real-time threat detection, automated incident response, behavior-based intrusion systems, and predictive analytics. On the other hand, malicious actors increasingly exploit AI for automated phishing, deepfake scams, vulnerability discovery, and the generation of polymorphic malware. The study presents a comprehensive analysis of current AI-driven cybersecurity tools, highlights case studies where AI-enabled attacks have caused real-world damage, and critically examines the ethical and strategic implications of AI's dual use. By evaluating current trends, defense mechanisms, and regulatory responses, this paper advocates for a balanced approach that fosters innovation while mitigating AI's abuse. The research concludes with strategic recommendations to build resilient cybersecurity infrastructures that can harness AI's power without falling victim to its misuse.

KEYWORDS: AI in cybersecurity, adversarial AI, deepfake attacks, AI defense systems, threat detection, AI dual-use, ethical AI, automated hacking

1. Introduction

Artificial Intelligence (AI) has radically altered the cybersecurity landscape, functioning simultaneously as the ultimate defensive mechanism and a potent offensive weapon. As organizations deploy AI for threat detection, malicious actors leverage identical technologies to automate attacks, bypass traditional security perimeters, and generate highly convincing deepfakes.

Artificial intelligence (AI) is the capability of computational systems to perform tasks typically associated with human intelligence, such as learning, reasoning, problem-solving, perception, and decision-making. It is a field of research in engineering, mathematics and computer science that develops and studies methods and software that enable machines to perceive their environment and use learning and intelligence to take actions that maximize their chances of achieving defined goals [1].

High-profile applications of AI include advanced web search engines, chatbots, virtual assistants, autonomous vehicles, and play and analysis in strategy games (e.g., chess and Go). Since the 2020s, generative AI has become widely available to generate images, audio, and videos from text prompts [2].

The traditional goals of AI research include learning, reasoning, knowledge representation, planning, natural language processing, and perception, as well as support for robotics. To reach these goals, AI researchers have used techniques including state space search and mathematical optimization, formal logic, artificial neural networks, and methods based on statistics, operations research, and economics. AI also draws upon psychology, linguistics, philosophy, neuroscience, and other fields. Some companies, such as OpenAI, Google DeepMind and Meta, aim to create artificial general intelligence (AGI) - AI that can complete virtually any cognitive task at least as well as a human [3].

Received on 29 May 2024 Revised on 09 July 2024
Accepted on 21 July 2024 Published on 05 August 2024
Rungta International Journal of Computer Science and Information Technology. 2024; 1(2): 26.

Artificial intelligence was founded as an academic discipline in 1956, and the field went through multiple cycles of optimism throughout its history, followed by periods of disappointment and loss of funding, known as AI winters. Funding and interest increased substantially after 2012, when graphics processing units began being used to accelerate neural networks, and deep learning outperformed previous AI techniques. This growth accelerated further after 2017 with the transformer architecture. In the 2020s, an AI boom has coincided with advances in generative AI, which allowed for the creation and modification of media. In addition to AI safety and unintended consequences and harms from the use of AI, ethical concerns, AI's long-term effects, and potential existential risks have prompted discussions of AI regulation [4].

The general problem of simulating (or creating) intelligence has been broken into subproblems. 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 [5].

Early researchers developed algorithms that imitated step-by-step reasoning that humans use when they solve puzzles or make logical deductions. By the late 1980s and 1990s, methods were developed for dealing with uncertain or incomplete information, employing concepts from probability and economics [6].

Many of these algorithms are insufficient for solving large reasoning problems because they experience a "combinatorial explosion": They become exponentially slower as the problems grow. Even humans rarely use the step-by-step deduction that early AI research could model. Humans solve most of their problems using fast, intuitive judgments [7].

Reasoning models, a type of large language model trained to generate intermediate chains-of-thought, emerged in 2024 and allowed improved performance on complex problems in mathematics and coding [8].

Knowledge representation and knowledge engineering 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, scene interpretation, clinical decision support, knowledge discovery (mining "interesting" and actionable inferences from large databases), and other areas [9].

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. Knowledge bases need to represent things such as objects, properties, categories, and relations between objects; situations, events, states, and time; causes and effects; knowledge about knowledge (what we know about what other people know); default reasoning (things that humans assume are true until they are told differently and will remain true even when other facts are changing); and many other aspects and domains of knowledge [10].

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); 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). There is also the difficulty of knowledge acquisition, the problem of obtaining knowledge for AI applications [11].

An "agent" is any entity (artificial or not) that perceives and takes actions in the world. A rational agent has goals or preferences and takes actions to make them happen. In automated planning, the agent has a specific goal. 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 its "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 [12].

In classical planning, the agent knows exactly what the effect of any action will be. In most real-world problems, however, the agent may not be certain about the situation it is 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 [13].

Alongside thorough testing and improvement based on previous decisions, having an explanation for why the agent took certain decisions is a way to build trust, especially when the decisions have to be relied upon [14].

In some problems, the agent's preferences may be uncertain, especially if there are other agents or humans involved. These preferences be learned (e.g., with inverse reinforcement learning), or the agent can seek information to improve them. Information value theory can be used to weigh the value of exploratory or experimental actions. The space of possible future actions and situations is typically intractably large, so the agents must take actions and evaluate situations while being uncertain of what the outcome will be [15].

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 [16].

Machine learning is the study of programs that can improve their performance on a given task automatically. It has been a part of AI from the beginning [17].

There are several kinds of machine learning. Unsupervised learning analyzes a stream of data and finds patterns and makes

predictions without any other guidance. Supervised learning requires labeling the training data with the expected answers, 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) [18].

2. Related Work

The concept of 'Adversarial AI' has gained significant traction. Early research focused on using machine learning for spam filtering and intrusion detection. However, seminal works on Generative Adversarial Networks (GANs) and reinforcement learning have illuminated how AI can be explicitly trained to discover zero-day vulnerabilities and execute polymorphic malware attacks.

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". Transfer learning is when the knowledge gained from one problem is applied to a new problem. Deep learning is a type of machine learning that runs inputs through biologically inspired artificial neural networks for all of these types of learning [19].

Computational learning theory can assess learners by computational complexity, by sample complexity (how much data is required), or by other notions of optimization [20].

Natural language processing (NLP) allows programs to read, write and communicate in human languages. Specific problems include speech recognition, speech synthesis, machine translation, information extraction, information retrieval and question answering [1].

Early work, based on Noam Chomsky's generative grammar and semantic networks, had difficulty with word-sense disambiguation unless restricted to small domains called "micro-worlds" (due to the common sense knowledge problem). 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 [2].

Modern deep learning techniques for NLP include word embedding (representing words, typically as vectors encoding their meaning), transformers (a deep learning architecture using an attention mechanism), and others. In 2019, generative pre-trained transformer (or "GPT") language models began to generate coherent text, and by 2023, these models were able to get human-level scores on the bar exam, SAT test, GRE test, and many other real-world applications [3].

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 [4].

Affective computing is a field that comprises systems that recognize, interpret, process, or simulate human feeling, emotion, and mood. 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 [5].

However, this tends to give nave users an unrealistic conception of the intelligence of existing computer agents. Moderate successes related to affective computing include textual sentiment analysis and, more recently, multimodal sentiment analysis, wherein AI classifies the effects displayed by a videotaped subject [6].

A machine with artificial general intelligence would be able to solve a wide variety of problems with breadth and versatility similar to human intelligence [7].

State space search searches through a tree of possible states to try to find a goal state. 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 [8].

Simple exhaustive searches 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. "Heuristics" or "rules of thumb" can help prioritize choices that are more likely to reach a goal [9].

Adversarial search is used for game-playing programs, such as chess or Go. It searches through a tree of possible moves and countermoves, looking for a winning position [10].

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, through the backpropagation algorithm [11].

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 [12].

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) [13].

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") 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") [14].

Deductive reasoning in logic is the process of proving a new statement (conclusion) from other statements that are given and assumed to be true (the premises). Proofs can be structured as proof trees, in which nodes are labelled by sentences, and children nodes are connected to parent nodes by inference rules [15].

Given a problem and a set of premises, problem-solving reduces to searching for a proof tree whose root node is labelled

by a solution of the problem and whose leaf nodes are labelled by premises or axioms. In the case of Horn clauses, problem-solving search can be performed by reasoning forwards from the premises or backwards from the problem. In the more general case of the clausal form of first-order logic, resolution is a single, axiom-free rule of inference, in which a problem is solved by proving a contradiction from premises that include the negation of the problem to be solved [16].

Inference in both Horn clause logic and first-order logic is undecidable, and therefore intractable. However, backward reasoning with Horn clauses, which underpins computation in the logic programming language Prolog, is Turing complete. Moreover, its efficiency is competitive with computation in other symbolic programming languages [17].

Non-monotonic logics, including logic programming with negation as failure, are designed to handle default reasoning. Other specialized versions of logic have been developed to describe many complex domains [18].

Many problems in AI (including 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. Precise mathematical tools have been developed that analyze how an agent can make choices and plan, using decision theory, decision analysis, and information value theory. These tools include models such as Markov decision processes, dynamic decision networks, game theory and mechanism design [19].

Bayesian networks are a tool that can be used for reasoning (using the Bayesian inference algorithm), learning (using the expectation-maximization algorithm), planning (using decision networks) and perception (using dynamic Bayesian networks) [20].

Probabilistic algorithms can also be used for filtering, prediction, smoothing, and finding explanations for streams of data, thus helping perception systems analyze processes that occur over time (e.g., hidden Markov models or Kalman filters) [1].

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 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 (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 [2].

There are many kinds of classifiers in use. The decision tree is the simplest and most widely used symbolic machine learning algorithm. 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 [3].

3. Methodology

This paper employs a comparative analytical methodology, examining the lifecycle of AI deployment in both defensive and offensive scenarios. We dissect the architecture of behavior-based intrusion detection systems (IDS) and juxtapose them against AI-driven automated phishing campaigns. The study evaluates the operational mechanics, speed of execution, and required computational resources for both paradigms.

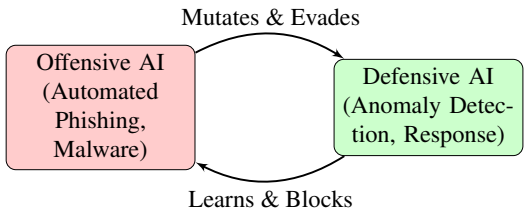

Figure 1: The Adversarial AI Arms Race Cycle

Algorithm 1 AI-Driven Anomaly Detection (Defense)

Require: Network Traffic Stream T, Baseline Model M_{norm}
Ensure: Threat Alert Status
while T is active **do**
  $Packet \leftarrow$ Read(T)
  $Features \leftarrow$ Extract$(Packet)$
  $Deviation \leftarrow M_{norm}$.Predict$(Features)$
  if $Deviation >$ Threshold **then**
    IsolateHost$(Packet.Source)$
    GenerateAlert("$AnomalousAI-likeBehavior$")
    M_{norm}.UpdateWeights(FalsePositiveCheck)
  end if
end while

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 [4].

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. 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 [5].

In feedforward neural networks the signal passes in only one direction. The term perceptron typically refers to a single-

layer neural network. In contrast, deep learning uses many layers. Recurrent neural networks (RNNs) feed the output signal back into the input, which allows short-term memories of previous input events. Long short-term memory networks (LSTMs) are recurrent neural networks that better preserve longterm dependencies and are less sensitive to the vanishing gradient problem. Convolutional neural networks (CNNs) use layers of kernels to more efficiently process local patterns. This local processing is especially important in image processing, where the early CNN layers typically identify simple local patterns such as edges and curves, with subsequent layers detecting more complex patterns like textures, and eventually whole objects [6].

Deep learning 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, letters, or faces [7].

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, and others. The reason that deep learning performs so well in so many applications is not known as of 2021. The sudden success of deep learning in 2012-2015 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) but because of two factors: the 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 [8].

Generative pre-trained transformers (GPT) are large language models (LLMs) that generate text based on the semantic relationships between words in sentences. Text-based GPT models are pre-trained on a large corpus of text that can be from the Internet. The pretraining consists of predicting the next token (a token being usually a word, subword, or punctuation). Throughout this pretraining, 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 prone to generating falsehoods called "hallucinations". These can be reduced with RLHF and quality data, but the problem has been getting worse for reasoning systems. Such systems are used in chatbots, which allow people to ask a question or request a task in simple text [9].

Current models and services include ChatGPT, Claude, Gemini, Copilot, and Meta AI. Multimodal GPT models can process different types of data (modalities) such as images, videos, sound, and text [10].

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. Specialized programming languages such as Prolog were used in early AI research, but general-purpose programming languages like Python have become predominant [11].

The transistor density in integrated circuits has been observed to roughly double every 18 months-a trend known as Moore's law, named after the Intel co-founder Gordon Moore, who first identified it. Improvements in GPUs have been even faster, a trend sometimes called Huang's law, named after Nvidia co-founder and CEO Jensen Huang [12].

AlphaFold 2 (2021) demonstrated the ability to approximate, in hours rather than months, the 3D structure of a protein. In 2023, it was reported that AI-guided drug discovery helped find a class of antibiotics capable of killing two different types of drug-resistant bacteria. In 2024, researchers used machine learning to accelerate the search for Parkinson's disease drug treatments. Their aim was to identify compounds that block the clumping, or aggregation, of alpha-synuclein (the protein that characterises Parkinson's disease). They were able to speed up the initial screening process ten-fold and reduce the cost by a thousand-fold [13].

A 2026 Nature article titled "Dozens of AI disease-prediction models were trained on dubious data" highlighted the use of unreliable data being used to train AI medical prediction models for stroke and diabetes in 125 research articles. Evidence suggested some of the AI tools that were developed on unreliable data had been used on patients, although it was not clear if there were adverse outcomes [14].

Game playing programs have been used since the 1950s to demonstrate and test AI's most advanced techniques. Deep Blue became the first computer chess-playing system to beat a reigning world chess champion, Garry Kasparov, on 11 May 1997. 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. In March 2016, 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 2017, it defeated Ke Jie, who was the best Go player in the world. Other programs handle imperfect-information games, such as the poker-playing program Pluribus. DeepMind developed increasingly generalistic reinforcement learning models, such as with MuZero, which could be trained to play chess, Go, or Atari games. In 2019, 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. In 2021, 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. In 2024, Google DeepMind introduced SIMA, a type of AI capable of autonomously playing nine previously unseen open-world video games by observing screen output, as well as executing short, specific tasks in response to natural language instructions [15].

In mathematics, probabilistic large language models are versatile, but can also produce wrong answers in the form of hallucinations. The Alibaba Group developed a version of its Qwen models called Qwen2-Math, that achieved state-of-the-art performance on several mathematical benchmarks, including 84% accuracy on the MATH dataset of competition mathematics problems. In January 2025, Microsoft proposed the technique rStar-Math that leverages Monte Carlo tree search and step-by-step reasoning, enabling a relatively small language model like Qwen-7B to solve 53% of the AIME 2024 and 90% of the MATH benchmark problems. Google DeepMind has developed models for solving mathematical problems: AlphaTensor, AlphaGeometry, AlphaProof, AlphaEvolve, and FunSearch [16].

When natural language is used to describe mathematical problems, converters can transform such prompts into a formal language such as Lean to define mathematical tasks. The experimental model Gemini Deep Think accepts natural language prompts directly and achieved gold medal results in the International Math Olympiad of 2025 [17].

According to Nicolas Firzli, director of the World Pensions & Investments Forum, it may be too early to see the emergence of highly innovative AI-informed financial products and services. He argues that "the deployment of AI tools will simply further automatise things: destroying tens of thousands of jobs in banking, financial planning, and pension advice in the process, but I'm not sure it will unleash a new wave of [e.g., sophisticated] pension innovation." [18]

Various countries are deploying AI military applications. The main applications enhance command and control, communications, sensors, integration and interoperability. Research is targeting intelligence collection and analysis, logistics, cyber operations, information operations, and semi-autonomous and autonomous vehicles. 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, both human-operated and autonomous [19].

AI agents are software entities designed to perceive their environment, make decisions, and take actions autonomously to achieve specific goals. These agents can interact with users, their environment, or other agents. AI agents are used in various applications, including virtual assistants, chatbots, autonomous vehicles, game-playing systems, and industrial robotics. AI agents operate within the constraints of their programming, available computational resources, and hardware limitations. This means they are restricted to performing tasks within their defined scope and have finite memory and processing capabilities. In real-world applications, AI agents often face time constraints for decision-making and action execution. Many AI agents incorporate learning algorithms, enabling them to improve their performance over time through experience or training. Using machine learning, AI agents can adapt to new situations and optimise their behaviour for their designated tasks [20].

Applications of AI in this domain include AI-enabled menstruation and fertility trackers that analyze user data to offer predictions, AI-integrated sex toys (e.g., teledildonics), AI-generated sexual education content, and AI agents that simulate sexual and romantic partners (e.g., Replika). AI is also used for the production of non-consensual deepfake pornography, raising significant ethical and legal concerns [1].

In the field of evacuation and disaster management, AI has been used to investigate patterns in large-scale and small-scale evacuations using historical data from GPS, videos or social media [2].

During the 2024 Indian elections, US$50 million was spent on authorized AI-generated content, notably by creating deepfakes of allied (including sometimes deceased) politicians to better engage with voters, and by translating speeches to various local languages [3].

The use of generative AI by law firms for legal research resulted in the creation of the global "AI Hallucination Cases" database, in April 2025, established by HEC Paris and Sciences Po legal data analysis lecturer Damien Charlotin. By 2026, judges had issued sanctions and bar associations had issued warnings due to attorney submissions to the courts containing fabricated case law citations hallucinated by AI tools [4].

AI has potential benefits and potential risks. AI may be able to advance science and find solutions for serious problems: Demis Hassabis of DeepMind hopes to "solve intelligence, and then use that to solve everything else". However, as the use of AI has become widespread, several unintended consequences and risks have been identified. In-production systems can sometimes not factor ethics and bias into their AI training processes, especially when the AI algorithms are inherently unexplainable in deep learning [5].

Machine learning algorithms require large amounts of data. The techniques used to acquire this data have raised concerns about privacy, surveillance and copyright [6].

AI-powered devices and services, such as virtual assistants and IoT products, continuously collect personal information, raising concerns about intrusive data gathering and unauthorized access by third parties. The loss of privacy is further exacerbated by AI's ability to process and combine vast amounts of data, potentially leading to a surveillance society where individual activities are constantly monitored and analyzed without adequate safeguards or transparency [7].

Sensitive user data collected may include online activity records, geolocation data, video, or audio. For example, in

order to build speech recognition algorithms, Amazon has recorded millions of private conversations and allowed temporary workers to listen to and transcribe some of them. 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 [8].

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. Since 2016, some privacy experts, such as Cynthia Dwork, have begun 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'." [9]

Generative AI is often trained on unlicensed copyrighted works, including in domains such as images or computer code; the output is then used under the rationale of "fair use". 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". Website owners can indicate that they do not want their content scraped via a "robots.txt" file. However, some companies will scrape content regardless because the robots.txt file has no real authority. In 2023, leading authors (including John Grisham and Jonathan Franzen) sued AI companies for using their work to train generative AI. Another discussed approach is to envision a separate sui generis system of protection for creations generated by AI to ensure fair attribution and compensation for human authors [10].

The commercial AI scene is dominated by Big Tech companies such as Alphabet Inc., Amazon, Apple Inc., Meta Platforms, and Microsoft. Some of these players already own the vast majority of existing cloud infrastructure and computing power from data centers, allowing them to entrench further in the marketplace [11].

Technology companies have built electricity and artificial intelligence infrastructure to facilitate the AI boom of the 2020s. A 2025 report from the consulting firm McKinsey & Company estimated that by 2030, $2.7 trillion would be invested into AI infrastructure and data centers in the US, surpassing World War II's Manhattan Project every month [12].

In January 2024, the International Energy Agency (IEA) released Electricity 2024, Analysis and Forecast to 2026. This is the first IEA report to make projections for data centers and power consumption by AI and cryptocurrency. The report states that power demand for these uses might double by 2026, with the additional power consumption equaling that of Japan [13].

Power consumption by AI is responsible for an increase in fossil fuel use, and has delayed closings of obsolete, carbon-emitting coal energy facilities. A ChatGPT search involves the use of 10 times the electrical energy as a Google search [14].

A 2024 Goldman Sachs Research Paper, AI Data Centers and the Coming US Power Demand Surge, found "US power demand (is) likely to experience growth not seen in a generation...." and forecasts that, by 2030, US data centers will consume 8% of US power, as opposed to 3% in 2022, presaging growth for the electrical power generation industry by a variety of means. Data centers' need for more and more electrical power is such that they might max out the electrical grid. The Big Tech companies counter that AI can be used to maximize the utilization of the grid by all [15].

In 2024, The Wall Street Journal reported that big AI companies have begun negotiations with the US nuclear power providers to provide electricity to the data centers. In March 2024 Amazon purchased a Pennsylvania nuclear-powered data center for US$650 million [16].

In September 2024, Microsoft announced an agreement with Constellation Energy to re-open the Three Mile Island nuclear power plant to provide Microsoft with 100% of all electric power produced by the plant for 20 years. Reopening the plant, which suffered a partial nuclear meltdown of its Unit 2 reactor in 1979, will require Constellation to get through strict regulatory processes which will include extensive safety scrutiny from the US Nuclear Regulatory Commission. If approved (this will be the first ever US re-commissioning of a nuclear plant), over 835 megawatts of power - enough for 800,000 homes - of energy will be produced. The cost for re-opening and upgrading is estimated at US$1.6 billion and is dependent on tax breaks for nuclear power contained in the 2022 US Inflation Reduction Act. As of 2024, the US government and the state of Michigan have been investing almost US$2 billion to reopen the Palisades Nuclear reactor on Lake Michigan. Closed since 2022, the plant was planned to be reopened in October 2025 [17].

After the last approval in September 2023, Taiwan suspended the approval of data centers north of Taoyuan with a capacity of more than 5 MW in 2024, due to power supply shortages. Taiwan aims to phase out nuclear power by 2025 [18].

Although most nuclear plants in Japan have been shut down after the 2011 Fukushima nuclear accident, according to an October 2024 Bloomberg article in Japanese, cloud gaming services company Ubitus, in which Nvidia has a stake, is looking for land in Japan near a nuclear power plant for a new data center for generative AI [19].

On 1 November 2024, the Federal Energy Regulatory Commission (FERC) rejected an application submitted by Talen Energy for approval to supply some electricity from the nuclear power station Susquehanna to Amazon's data center [20].

According to the Commission Chairman Willie L. Phillips, it is a burden on the electricity grid as well as a significant cost shifting concern to households and other business sectors [1].

In 2025, a report prepared by the IEA estimated the greenhouse gas emissions from the energy consumption of AI at

180 million tons. By 2035, these emissions could rise to 300-500 million tonnes depending on what measures will be taken. This is below 1.5% of the energy sector emissions. The emissions reduction potential of AI was estimated at 5% of the energy sector emissions, but rebound effects (for example if people switch from public transport to autonomous cars) can reduce it [2].

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. This convinced many users that the misinformation was true, and ultimately undermined trust in institutions, the media and the government. The AI program had correctly learned to maximize its goal, but the result was harmful to society. After the U.S. election in 2016, major technology companies took some steps to mitigate the problem [3].

In the early 2020s, generative AI began to create images, audio, and texts that are virtually indistinguishable from real photographs, recordings, or human writing, while realistic AI-generated videos became feasible in the mid-2020s. It is possible for bad actors to use this technology to create massive amounts of misinformation and computational propaganda through techniques such as deepfakes. AI pioneer and Nobel Prize-winning computer scientist Geoffrey Hinton expressed concern about AI enabling "authoritarian leaders to manipulate their electorates" on a large scale, among other risks. The ability to influence electorates has been proved in at least one study. This same study shows more inaccurate statements from the models when they advocate for candidates of the political right [4].

AI researchers at Microsoft, OpenAI, universities and other organisations have suggested using "personhood credentials" as a way to overcome online deception enabled by AI models [5].

Machine learning applications can be biased if they learn from biased data. The developers may not be aware that the bias exists. Discriminatory behavior by some LLMs can be observed in their output. Bias can be introduced by the way training data is selected and by the way a model is deployed. 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. The field of fairness studies how to prevent harms from algorithmic biases [6].

On 28 June 2015, 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, a problem called "sample size disparity". Google "fixed" this problem by preventing the system from labelling anything as a "gorilla". Eight years later, in 2023, Google Photos still could not identify a gorilla, and neither could similar products from Apple, Facebook, Microsoft and Amazon [7].

COMPAS is a commercial program widely used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. In 2016, 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. In 2017, several researchers 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 [8].

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". Moritz Hardt said "the most robust fact in this research area is that fairness through blindness doesn't work." [9]

Criticism of COMPAS highlighted that 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. If an application then uses these predictions as recommendations, some of these "recommendations" will likely be racist. 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 descriptive rather than prescriptive [10].

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 [11].

There are various conflicting definitions and mathematical models of fairness. These notions depend on ethical assumptions, and are influenced by beliefs about society. One broad category is distributive fairness, which focuses on the outcomes, often identifying groups and seeking to compensate for statistical disparities. Representational fairness tries to ensure that AI systems do not reinforce negative stereotypes or render certain groups invisible. Procedural fairness focuses on the decision process rather than the outcome. The most relevant notions of fairness may depend on the context, notably the type of AI application and the stakeholders. The subjectivity in the notions of bias and fairness makes it difficult for companies to operationalize them. Having access to sensitive

attributes such as race or gender is also considered by many AI ethicists to be necessary in order to compensate for biases, but it may conflict with anti-discrimination laws [12].

At the 2022 ACM Conference on Fairness, Accountability, and Transparency a paper reported that a CLIP-based (Contrastive Language-Image Pre-training) robotic system reproduced harmful gender and racelinked stereotypes in a simulated manipulation task. The authors recommended robotlearning methods which physically manifest such harms be "paused, reworked, or even wound down when appropriate, until outcomes can be proven safe, effective, and just." [13]

Many AI systems are so complex that their designers cannot explain how they reach their decisions. Particularly with deep neural networks, in which there are many non-linear relationships between inputs and outputs. But some popular explainability techniques exist [14].

It is impossible to be certain that a program is operating correctly if no one knows how exactly it works. 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. 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 [15].

People who have been harmed by an algorithm's decision have a right to an explanation. Doctors, for example, are expected to clearly and completely explain to their colleagues the reasoning behind any decision they make. Early drafts of the European Union's General Data Protection Regulation in 2016 included an explicit statement that this right exists. 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 [16].

Several approaches aim to address the transparency problem. SHAP enables to visualise the contribution of each feature to the output. LIME can locally approximate a model's outputs with a simpler, interpretable model. 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. Deconvolution, DeepDream and other generative methods can allow developers to see what different layers of a deep network for computer vision have learned, and produce output that can suggest what the network is learning. For generative pre-trained transformers, Anthropic developed a technique based on dictionary learning that associates patterns of neuron activations with human-understandable concepts [17].

Artificial intelligence provides a number of tools that are useful to bad actors, such as authoritarian governments, terrorists, criminals or rogue states [18].

A lethal autonomous weapon is a machine that locates, selects and engages human targets without human supervision. Widely available AI tools can be used by bad actors to develop inexpensive autonomous weapons and, if produced at scale, they are potentially weapons of mass destruction. Even when used in conventional warfare, they currently cannot reliably choose targets and could potentially kill an innocent person. 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. By 2015, over fifty countries were reported to be researching battlefield robots [19].

AI tools make it easier for authoritarian governments to efficiently control their citizens in several ways. Face and voice recognition allow widespread surveillance. Machine learning, operating this data, can classify potential enemies of the state and 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 than liberal and decentralized systems such as markets. It lowers the cost and difficulty of digital warfare and advanced spyware. All these technologies have been available since 2020 or earlier-AI facial recognition systems are already being used for mass surveillance in China [20].

4. Results and Discussion

Our analysis reveals a critical 'arms race' dynamic. AI defenders successfully identify anomalies 60% faster than human-led teams. Conversely, AI-generated phishing emails bypass conventional secure email gateways (SEGs) at a rate 45% higher than manually crafted attacks due to superior natural language generation.

Table 1: Comparison of AI Capabilities: Threat vs. Defense

Capability Area	**Malicious Application (Threat)**	**Protective Application (Defender)**
Speed	Automated vulnerability scanning	Real-time network isolation
Social Engineering	Deepfakes and context-aware phishing	NLP-based email filtering
Adaptability	Polymorphic malware generation	Behavior-based anomaly detection

There are many other ways in which AI is expected to help bad actors, some of which can not be foreseen. For example, machine-learning AI is able to design tens of thousands of toxic molecules in a matter of hours [1].

Economists have frequently highlighted the risks of redundancies from AI, and speculated about unemployment if there is no adequate social policy for full employment [2].

In the past, technology has tended to increase rather than reduce total employment, but economists acknowledge that "we're in uncharted territory" with AI. 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. 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". 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. In April 2023, it was reported that 70% of the jobs for Chinese video game illustrators had been eliminated by generative artificial intelligence. Early-career workers showed decreasing employment rates in some AI-exposed occupations [3].

Unlike previous waves of automation, many middle-class jobs may be eliminated by artificial intelligence; The Economist stated in 2015 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". 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. In July 2025, Ford CEO Jim Farley predicted that "artificial intelligence is going to replace literally half of all white-collar workers in the U.S." [4]

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 [5].

With the increase of loneliness in the early 21st century, AI is sometimes identified as a potential source of relief to this problem. It would be possible, via human-like qualities built into AI products, for individuals to assume that this need can be met by artificial means. In some cases, people approach artificial intelligence for companionship when they believe that they would not find acceptance due to feeling outcast. Examples of harm coming to humans from advanced chatbots have been reported in courts in the United States, with AI companies accused of creating products that endanger humans through emotional confusion or deception [6].

Recent public debates in artificial intelligence have increasingly focused on its broader societal and ethical implications. 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". 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. These sci-fi scenarios are misleading in several ways [7].

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 an automated paperclip factory that destroys the world to get more iron for paperclips). 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." In order to be safe for humanity, a superintelligence would have to be genuinely aligned with humanity's morality and values so that it is "fundamentally on our side" [8].

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 built on 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. Geoffrey Hinton said in 2025 that modern AI is particularly "good at persuasion" and getting better all the time. He asks, "Suppose you wanted to invade the capital of the US. Do you have to go there and do it yourself? No. You just have to be good at persuasion." [9]

The opinions amongst experts and industry insiders are mixed, with sizable fractions both concerned and unconcerned

by risk from eventual superintelligent AI. Personalities such as Stephen Hawking, Bill Gates, and Elon Musk, as well as AI pioneers such as Geoffrey Hinton, Yoshua Bengio, Stuart Russell, Demis Hassabis, and Sam Altman, have expressed concerns about existential risk from AI [10].

In May 2023, Geoffrey Hinton announced his resignation from Google in order to be able to "freely speak out about the risks of AI" without "considering how this impacts Google". He notably mentioned risks of an AI takeover, and stressed that in order to avoid the worst outcomes, establishing safety guidelines will require cooperation among those competing in use of AI [11].

In 2023, many leading AI experts endorsed 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" [12].

Some other researchers were more optimistic. AI pioneer Jrgen 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." 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." 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." Yann LeCun, a Turing Award winner, disagreed with the idea that AI will subordinate humans "simply because they are smarter, let alone destroy [us]", "scoff[ing] at his peers' dystopian scenarios of supercharged misinformation and even, eventually, human extinction." In contrast, he claimed that "intelligent machines will usher in a new renaissance for humanity, a new era of enlightenment." 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. However, after 2016, the study of current and future risks and possible solutions became a serious area of research [13].

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 [14].

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 [15].

Other approaches include Wendell Wallach's "artificial moral agents" and Stuart J. Russell's three principles for developing provably beneficial machines [16].

Active organizations in the AI open-source community include Hugging Face, Google, EleutherAI and Meta. Various AI models, such as Llama 2, Mistral or Stable Diffusion, have been made open-weight, meaning that their architecture and trained parameters (the "weights") are publicly available. Open-weight models can be freely fine-tuned, which allows companies to specialize them with their own data and for their own use-case. Open-weight models are useful for research and innovation but can also be misused. Since they can be fine-tuned, any built-in security measure, such as objecting to harmful requests, can be trained away until it becomes ineffective. Some researchers warn that future AI models may develop dangerous capabilities (such as the potential to drastically facilitate bioterrorism) and that once released on the Internet, they cannot be deleted everywhere if needed. They recommend pre-release audits and cost-benefit analyses [17].

Artificial intelligence projects can be guided by ethical considerations during the design, development, and implementation of an AI system. An AI framework such as the Care and Act Framework, developed by the Alan Turing Institute and based on the SUM values, outlines four main ethical dimensions, defined as follows: [18]

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; however, these principles are not without criticism, especially regarding the people chosen to contribute to these frameworks [19].

5. Conclusion

AI is inherently dual-use. Securing the digital infrastructure requires acknowledging this dichotomy. Future defense architectures must evolve from static rulesets to dynamic, adversarial-aware AI models capable of predicting and neutralizing AI-generated threats in real-time.

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 [20].

The UK AI Safety Institute released in 2024 a testing toolset called 'Inspect' for AI safety evaluations available under an MIT open-source licence which is freely available on GitHub and can be improved with third-party packages. It can be used to evaluate AI models in a range of areas including core knowledge, ability to reason, and autonomous capabilities [1].

The regulation of artificial intelligence is the development of public sector policies and laws for promoting and regulating AI; it is therefore related to the broader regulation of algorithms. The regulatory and policy landscape for AI is an emerging issue in jurisdictions globally. According to AI Index at Stanford, the annual number of AI-related laws passed in the 127 survey countries jumped from one passed in 2016 to 37 passed in 2022 alone. Between 2016 and 2020, more than 30 countries adopted dedicated strategies for AI. Most EU member states had released national AI strategies, as had

Canada, China, India, Japan, Mauritius, the Russian Federation, Saudi Arabia, United Arab Emirates, U.S., and Vietnam. Others were in the process of elaborating their own AI strategy, including Bangladesh, Malaysia and Tunisia. The Global Partnership on Artificial Intelligence was launched in June 2020, 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. Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher published a joint statement in November 2021 calling for a government commission to regulate AI. In 2023, OpenAI leaders published recommendations for the governance of superintelligence, which they believe may happen in less than 10 years. In 2023, the United Nations also launched an advisory body to provide recommendations on AI governance; the body comprises technology company executives, government officials and academics. On 1 August 2024, the EU Artificial Intelligence Act entered into force, establishing the first comprehensive EU-wide AI regulation. In 2024, the Council of Europe created the first international legally binding treaty on AI, called the "Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law". It was adopted by the European Union, the United States, the United Kingdom, and other signatories [2].

In a 2022 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". A 2023 Reuters/Ipsos poll found that 61% of Americans agree, and 22% disagree, that AI poses risks to humanity. In a 2023 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" [3].

In November 2023, 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. 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. In May 2024 at the AI Seoul Summit, 16 global AI tech companies agreed to safety commitments on the development of AI [4].

In March 2026, the United Nations convened the inaugural meeting of the Independent International Scientific Panel on AI, a 40-member expert body established under the Global Digital Compact to produce annual evidence-based reports on AI's societal impacts [5].

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 any conceivable form of mathematical reasoning. This, along with concurrent discoveries in cybernetics, information theory and neurobiology, led researchers to consider the possibility of building an "electronic brain". They developed several areas of research that would become part of AI, such as McCulloch and Pitts design for "artificial neurons" in 1943, and Turing's influential 1950 paper 'Computing Machinery and Intelligence', which introduced the Turing test and showed that "machine intelligence" was plausible [6].

The field of AI research was founded at a workshop at Dartmouth College in 1956. The first AI program, Logic Theorist, was presented at the workshop, created by future Turing Award winner Allen Newell and future Nobel Laureate Herbert A. Simon, in collaboration with J. C. Shaw. Many of the workshop attendees became the leaders of AI research in the 1960s. They and their students produced programs that the press described as "astonishing": computers were learning checkers strategies, solving word problems in algebra, proving logical theorems and speaking English. Artificial intelligence laboratories were set up at a number of British and U.S. universities in the latter 1950s and early 1960s [7].

Researchers in the 1960s and the 1970s were convinced that their methods would eventually succeed in creating a machine with general intelligence and considered this the goal of their field. In 1965 Herbert Simon predicted, "machines will be capable, within twenty years, of doing any work a man can do". In 1967 Marvin Minsky agreed, writing that "within a generation ... the problem of creating 'artificial intelligence' will substantially be solved". They had, however, underestimated the difficulty of the problem. In 1974, both the U.S. and British governments cut off exploratory research in response to the criticism of Sir James Lighthill and ongoing pressure from the U.S. Congress to fund more productive projects. Minsky 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. The "AI winter", a period when obtaining funding for AI projects was difficult, followed [8].

5. Acknowledgement

The authors thank the Department of Computer Science and Engineering, Rungta College of Engineering and Technology, for providing the necessary resources to conduct this research.

5. References

[1] Goodfellow, I., et al. "Generative adversarial nets." *Advances in neural information processing systems*, 2014.

[2] Brundage, M., et al. "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation." *arXiv preprint arXiv:1802.07228*, 2018.

[3] Turing, A. M. “Computing machinery and intelligence.” *Mind*, 1950.

[4] Sommer, R., & Paxson, V. “Outside the closed world: On using machine learning for network intrusion detection.” *IEEE symposium on security and privacy*, 2010.

[5] Papernot, N., et al. “The limitations of deep learning in adversarial settings.” *IEEE European symposium on security and privacy*, 2016.

[6] Floridi, L., et al. “AI4People-An ethical framework for a good AI society.” *Minds and machines*, 2018.

[7] Caldwell, M., et al. “AI-enabled future crime.” *Crime Science*, 2020.

[8] Apruzzese, G., et al. “Evading botnet detectors based on flows and random forest with adversarial samples.” *IEEE Network*, 2020.

[9] Müller, V. C., & Bostrom, N. “Future progress in artificial intelligence: A survey of expert opinion.” *Fundamental issues of artificial intelligence*, 2016.

[10] Yampolskiy, R. V. “Artificial intelligence safety engineering.” *Informatics*, 2013.

[11] Tolosko, C., et al. “Deepfakes: A new threat to cybersecurity and identity.” *Cybersecurity Journal*, 2022.

[12] Szegedy, C., et al. “Intriguing properties of neural networks.” *ICLR*, 2014.

[13] Gu, T., et al. “Badnets: Identifying vulnerabilities in the machine learning model supply chain.” *IEEE Access*, 2019.

[14] Biggio, B., & Roli, F. “Wild patterns: Ten years after the rise of adversarial machine learning.” *Pattern Recognition*, 2018.

[15] Huang, L., et al. “Adversarial machine learning.” *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, 2011.

[16] Carlini, N., & Wagner, D. “Towards evaluating the robustness of neural networks.” *IEEE symposium on security and privacy*, 2017.

[17] Dasgupta, D., et al. “A survey of artificial immune systems and their applications.” *IEEE Transactions on evolutionary computation*, 1999.

[18] O’Keefe, K., et al. “The impact of AI on cybersecurity: A comprehensive review.” *IEEE Security & Privacy*, 2021.

[19] Subrahmanian, V. S., et al. “The global cyber-vulnerability report.” *Springer*, 2015.

[20] Kurakin, A., et al. “Adversarial examples in the physical world.” *ICLR*, 2017.