Review

Navigating Artificial General Intelligence (AGI): Societal Implications, Ethical Considerations, and Governance Strategies

Dileesh Chandra Bikkasani *

Posted Date: 19 July 2024

doi: 10.20944/preprints202407.1573.v1

Keywords: Artificial General Intelligence (AGI), AGI Ethics, AGI Governance, Security, Societal Impact

Review

Navigating Artificial General Intelligence (AGI): Societal Implications, Ethical Considerations, and Governance Strategies

Dileesh Chandra Bikkasani

University of Bridgeport, Connecticut, USA, contact: dbikkasa@my.bridgeport.edu

Abstract: Artificial General Intelligence (AGI) represents a pivotal advancement in AI with far-reaching implications across technological, ethical, and societal domains. This paper addresses the following: (1) an in-depth assessment of AGI's transformative potential across different sectors and its multifaceted implications, including significant financial impacts like workforce disruption, income inequality, productivity gains, and potential systemic risks; (2) an examination of critical ethical considerations, including transparency and accountability, complex ethical dilemmas and societal impact; (3) a detailed analysis of privacy, legal and policy implications, particularly in intellectual property and liability, and (4) a proposed governance framework to ensure responsible AGI development and deployment. Additionally, the paper explores and addresses AGI's political implications, including national security and potential misuse. By analyzing and considering computer science, philosophy, economics, and policy perspectives, we offer a multidisciplinary view of AGI's challenges and opportunities, advocating for proactive measures to align AGI development with human values and societal interests.

Keywords: Artificial General Intelligence (AGI)**;** AGI ethics**;** AGI governance**;** security**;** societal impact

1. Introduction

"Let us define an ultra-intelligent machine as one that could surpass human capabilities in all domains. Since "designing a machine" is one of those domains, it becomes a cycle of self-improvement called the Intelligence Explosion. The human who oversaw all this would be left far behind. Making AGI the final invention we will ever have made" (Good 1966).

Humans dominate the earth mainly due to our cognitive capabilities, such as language, reasoning, social interactions, energy, and tool usage. The development and expansion of the human brain has been a gradual process spanning millions of years (Defelipe 2011). AI has evolved at an unprecedented rate compared to the human brain due to technological advancements and increased computational power. The AI development landscape took a significant turn when Geoffrey Hinton introduced deep belief nets, paving the way for deep learning and developing many algorithms, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Generative Adversarial Networks (GAN) (Shao et al. 2022). A critical checkpoint in the path to AGI was the recent breakthrough in the field of Natural Language Processing (NLP), with the Large Language Models (LLMs) utilizing the transformer architecture and attention mechanism (Vaswani et al. 2017), a neural network was able to predict the next word in a sentence when trained on massive corpora of text.

Initially, AI researchers focused on "narrow AI," systems designed for specific tasks, such as games and pattern recognition, due to the complexity of achieving general intelligence. However, with the emergence of LLMs capable of generating human-like text and reasoning, concerns have grown about the societal impacts of AGI. As AI becomes good at language and reasoning, it is crucial to understand the implications for society, including job displacement, privacy invasions, and ethical challenges.

 doi:10.20944/preprints202407.1573.v1

The journey towards AGI requires a multidisciplinary approach, engaging academia, government, industry experts, and civil society to navigate the vast landscape of intelligent systems. There is also an existential risk associated with the development of AGI, including the possibility of an Artificial Superintelligence (ASI) that could pose an existential threat to humanity if not managed responsibly. On the one hand, AGI can change how we live by enhancing our lives. On the other hand, it raises ethical and existential concerns, such as the potential for job displacement, privacy issues, and the risk of creating systems that could surpass human control (Bostrom 2014).

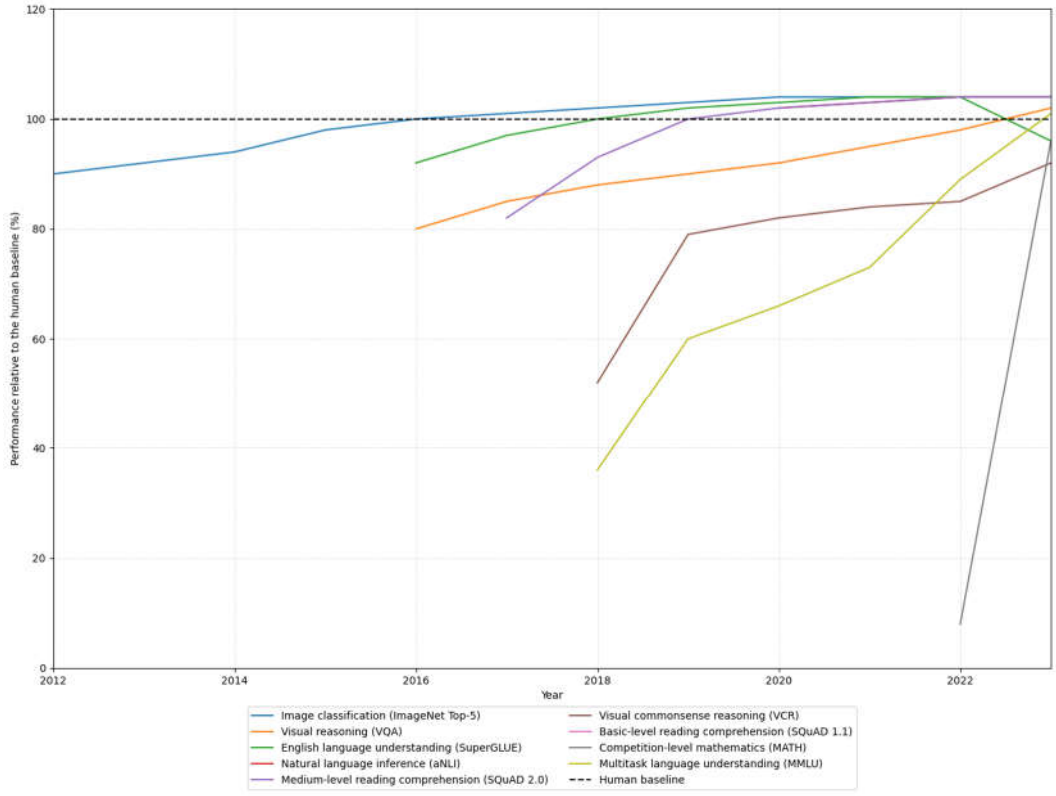
The governance of AGI presents a complex challenge, requiring revisiting the current regulatory frameworks and innovative frameworks to oversee development and deployment. Transparency, accountability, and ethical considerations ensure that AGI serves our best interests without compromising privacy and security. New proposals for ethical guidelines, oversight boards, and regulatory agencies are emerging to steer AGI development in a responsible direction.

As AGI moves from a theoretical possibility to a practical reality, it is crucial to consider the need for thoughtful governance, ethical oversight, and global collaboration. By encouraging a dialogue that engages stakeholders across different industries and prioritizes human values, we can harness the power of AGI while mitigating its risks, ensuring a future where humans are complemented by the system rather than a source of uncertainty and instability.

2. Current State of AGI

The path towards AGI has seen significant progress in recent years due to breakthroughs in machine learning, increased computational power, and the vast amounts of data collected, see Figure 1. Experts predict a technological 'singularity' by 2045 (Brynjolfsson et al. 2017).

A significant milestone in the pursuit of AGI was the AlphaGo program developed by DeepMind, which combined deep neural networks with techniques like Monte Carlo tree search and reinforcement learning to master the decision-making processes for the complex game of Go. In 2016, AlphaGo defeated Lee Sedol, one of the top Go players in the world, in a historic 4-1 victory (Silver et al. 2017). This demonstrates that through self-learning and playing against itself, AI can improve at a game known for its vast complexity, intuitive elements, and combinatorial challenges, which were traditionally thought to be mastered only by humans.



 doi:10.20944/preprints202407.1573.v1

Figure 1. Timeline showing AI's performance surpassing human performance on several benchmarks, including some in image classification, visual reasoning, and English understanding tasks. Source: AI index, 2024.

Machine learning algorithms have primarily driven recent advancements in AGI. The current approach to AGI research spans different methodologies. One notable approach is whole brain emulation (WBE), where researchers in Europe are working on the Human Brain Project, which combines symbolic reasoning with deep learning capabilities to enhance AGI (Marcus 2018). This approach bridges the gap between symbolic AI, known for logical reasoning, and the statistical strengths of deep learning models (Battaglia et al. 2018). Researchers from Google's DeepMind and Harvard trained a virtual agent using deep reinforcement learning and built a biomechanically realistic simulation of a rat to mimic actual rat behavior. This work advances our understanding of how the brain implements motor control and possibly allows us to model complex behaviors (Aldarondo et al. 2024).

With recent technological advancements, major tech companies like Google, Microsoft, OpenAI, and Anthropic are investing heavily in developing LLMs and AI systems that are multimodal and capable of perceiving, comprehending, and interacting with the world more like humans. Recent milestones include models such as OpenAI's GPT-3 (2020) and GPT-4 (2024), which show remarkable capabilities in natural language understanding and generation across different domains. Google's AlphaCode 2 demonstrated the ability of an AI system to compete with humans in the competitive programming space. With the integration of LLMs and training on 30+ million lines of code, this model could be placed in the 85th percentile (AlphaCode 2 technical report 2024). Another company, Anthropic, explored constitutional AI by instilling beneficial values into AI systems to increase intelligence across multiple domains (Bai et al. 2022).

Another focus for researchers is continual learning, which addresses the challenges of AGI and its ability to learn and retain knowledge over extended periods without forgetting previously learned information. Research in this area explores mechanisms such as elastic weight consolidation and synaptic intelligence, which allow models to learn new tasks while retaining knowledge from past experiences (Chaudhry et al. 2019).

In model architectures, developing transformer-based models customized for AGI tasks, such as integrating memory-augmented networks with transformers, enhances the ability to perform complex reasoning and decision-making (Rae et al. 2019). These models aim to emulate human reasoning processes by scaling cognitive capabilities and processing and manipulating information more effectively.

Furthermore, the generative capabilities of LLMs and their integration with other modalities, such as vision and robotics, have led to the development of new types of AI systems that can comprehensively interact with and perceive the world. Researchers are exploring hybrid models that combine rule-based systems and neural networks with memory to create cognitive architectures capable of reasoning and learning like humans. Another approach is transfer learning, which allows AI systems to transfer knowledge from one domain to another. Researchers from Microsoft, Google, MIT, and Oxford developed DenseAV, an AI algorithm that learns the meaning of words and sounds by watching videos, potentially advancing our understanding of how language and visual learning interact. This could lead to a human-like learning experience for AI systems (Hamilton et al. 2024).

In summary, the field of AGI is witnessing significant progress on multiple fronts, from WBE to advances in continual learning and transformer-based architectures. These developments are rapidly bridging the gap between AI and AGI systems. These advancements necessitate addressing the ethical considerations for their development and deployment in society, including the challenges surrounding governance, value alignment, and reliability, which must be resolved before taking the monumental step toward achieving AGI.

3. Implications for the Economy

 doi:10.20944/preprints202407.1573.v1

Automation already plays a vital role in our daily lives, and the widespread adoption of AI and automation will likely have far-reaching implications for the economy. A critical question is whether AI and its automation capabilities would complement or replace the human workforce. It depends on the type of work the workers are performing, and there is a great chance it might create additional opportunities and a "*reinstatement effect.*" (Acemoglu and Restrepo 2019). Although advancements in this space promise an increase in efficiency and output, they also raise concerns about potential job displacements.

Historically, technological advancements have led to new job opportunities and industries. As AI and automation technologies evolve, they may also give rise to new job roles and skill requirements. For instance, developing and maintaining AI systems will require a skilled workforce in data science, machine learning, and software engineering. According to the World Economic Forum, 97 million new job roles may be created due to the adoption of such technologies (World Economic Forum 2020).

However, one of the primary economic impacts of achieving AGI is the potential for job cuts, particularly in industries like manufacturing, data entry, customer service, and accounting, where many routine tasks can be automated. A study by McKinsey stated that automation would be responsible for replacing up to 800 million jobs by 2030 (Manyika et al. 2017). Industries such as manufacturing, transportation, and specific administrative roles may experience significant job disruptions as AI systems become more capable of performing traditional tasks done by human workers. For instance, self-driving vehicles and automated logistics systems could displace millions of truck drivers and delivery workers (Autor 2015).

As the development and control of such technologies lie in the hands of higher-skilled workers, it might lead to income inequality. A report by the International Monetary Fund states, "If AI significantly complements higher-income workers, it may lead to a disproportionate increase in their labor income." (Cazzaniga et al. 2024) which could destabilize economies. Addressing these challenges would require policy measures, including progressive taxation, universal basic income, and social safety nets. Promoting inclusive growth through investments in education and healthcare can ensure that the benefits from AGI can be broadly shared across society (OECD 2019).

Human capital is a crucial aspect of any economy. A typical timeline to develop a human worker, including education, is roughly 30 years (Bostrom 2014), depending on the expertise required for specific industries. This process requires significant investment in education and skill development. Unlike an AI, whose training time depends on the number of resources available, training an LLM, like the GPT-3 model, would take approximately 355 years on a single "*Graphic Processor Unit*" (GPU) (Baji 2017). In contrast, it could take around 34 days to train using massive clusters of GPUs and parallel processing (Narayanan et al. 2021). However, the training is a one-time process, and the skills could transfer across different domains, making it much cheaper to deploy new AI agents.

Another concern about AGI handling the financial markets is the inherent "*systemic risk.*" Systemic risk in finance refers to the risk of failure of the entire economic system, which arises from the interconnected nature of securities, where the failure of one system can cause a cascading effect on the whole system. An unconstrained AGI system tasked with maximizing profits without proper constraints could cause more significant damage than the 2010 flash crash, where a high-frequency trading algorithm rapidly sold S&P 500 E-mini futures contracts, causing stock market indices to drop up to 9% intraday (Staffs of CFTC and SEC 2010). The full consequences of an unconstrained profit-maximizing AGI system remain unknown.

Given these potential impacts, policymakers face a challenging environment in which to foster innovation while mitigating its economic risks. Some possible policy considerations could be implementing robust safety and ethics regulations, developing AGI-focused antitrust measures to prevent monopoly over markets, and creating retraining programs for displaced workers. As AGI development progresses, addressing these challenges proactively through thoughtful policy-making and inclusive dialogue is crucial.

4. Implications for Energy and Climate

 doi:10.20944/preprints202407.1573.v1

The development of AGI faces significant challenges in terms of energy consumption and sustainability. The environmental and ecological impacts are often overlooked when it comes to the advancements in this space. The computational power required to sustain AI models doubles every 100 days (Zhu et al. 2023). Increasing the capacity of a model by tenfold can result in a 10,000-fold rise in power demand. As AI systems become more advanced, their computational demands for training and running the system also increase, refer to Figure 2. Initiatives such as the Global Alliance on Artificial Intelligence for Industry and Manufacturing (AIM-Global) by the United Nations Industrial Development Organization (UNIDO) highlight the importance of aligning AI advancements with global sustainability goals, particularly in mitigating the environmental impacts associated with AI and AGI technologies (UNIDO 2023).

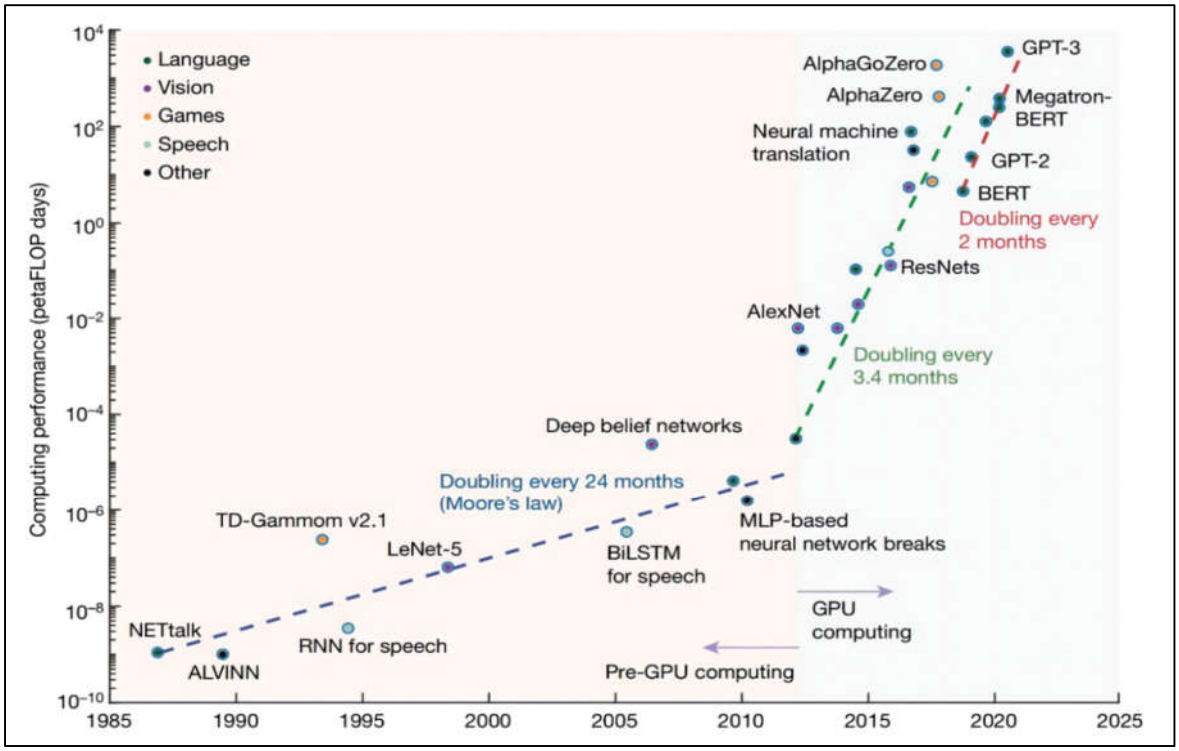


Figure 2. Over the past decade, growth in computing power demands substantially outpacing macro trends (Mehonic and Kenyon 2022).

The two phases of energy consumption for AI systems are training and inference, with training consuming around 20% and inference consuming 80% of the resources. The energy demand for Natural Language Processing (NLP) tasks is exceptionally high, as the models have to be trained on vast datasets. Researchers estimated that training GPT-3 would have consumed around 1300MWh (Patterson et al. 2021). In comparison, GPT-4 is estimated to have consumed 51,772- 62,318 MWh of energy, which is roughly equivalent to the monthly output of a nuclear power plant.

Additionally, the AI models have a significant carbon footprint, with some transformer neural networks emitting over 626,000 lbs. of CO2 (Kurshan 2023). Techniques such as power-capping the GPUs would reduce energy usage by 15% and only a marginal increase in computational time (McDonald et al. 2022). Another promising avenue is the development of AI-specific energy-efficient hardware designed for workloads. Companies like Nvidia, Google, and Intel are investing in AI chips and tensor processing units (TPUs) that deliver high performance while consuming less power than CPUs and GPUs (Sze et al. 2020). Distributed and federated learning are also being considered to distribute the computing load across multiple devices or edge nodes, reducing the energy demands on centralized data centers (Konečný et al. 2016).

 doi:10.20944/preprints202407.1573.v1

If current language models require such immense amounts of energy and computational power for training and inference, developing AGI would necessitate far more resources, leading to even more significant environmental impacts on society.

The pursuit of AGI would demand orders of magnitude more computational resources than current narrow AI models. As the AGI systems increase in scale and complexity, the energy requirement and carbon footprint would escalate exponentially, potentially straining existing energy infrastructure and exacerbating climate change concerns. Addressing sustainability, energy efficiency, and their challenges will be crucial to mitigate its societal and ecological consequences (Ammanath 2024).

Shifting towards renewable energy sources and energy-efficient computing infrastructure is essential to minimize the associated environmental impact. Leveraging renewable resources like solar, wind, and hydroelectric power can significantly reduce the strain on current infrastructure and the carbon footprint, aligning with global efforts to combat climate change.

5. Ethical Implications

Until the advent of machine learning, machines were only relied on to execute a programmed set of instructions. However, with the development of AI and ML systems, the decision-making framework is gradually shifting towards them. Such a transition raises questions about the ethics involved in their decision-making processes. For instance, driverless cars make decisions based on sensor information and the data used to train the algorithms, such as miles driven, driving patterns, and weather conditions. Discussing the ethical implications of decisions made by AGI systems is essential.

5.1. Transparency and Accountability

The development and research leading to AGI systems must be transparent, and the companies involved should be held accountable for the outcomes of such systems. Transparency helps identify algorithmic biases, instilling confidence in AI's ethical development and use. However, scrutinizing companies and algorithms is challenging due to the *trade secret* laws and regulations that protect them (Donovan et al. 2018). Despite these challenges, establishing a common framework is essential to maintain ethical standards.

5.2. Ethical Dilemmas

Ethical dilemmas, such as *The Trolley Problem*, highlight the challenges of AI decision-making in morally ambiguous situations. It presents a scenario where an autonomous vehicle must choose between two courses of action, each resulting in different casualties. For instance, given two choices, a car has to choose between a little girl and an older adult. These ethical dilemmas are critical in self-driving vehicles, which must make split-second decisions involving human lives. There are a few instances where the lack of opacity of the self-driving system has led to complex legal and liability issues (Griggs and Wakabayashi 2018). Where a self-driving Uber killed a person, and the driver was held responsible. The decision-making process and its lack of transparency can lead to legal and liability concerns, as seen in the case of self-driving vehicles that have been involved in fatal accidents. Establishing robust governance frameworks is crucial to address such incidents (Bird et al. 2020). Beyond this scenario, there are diverse ethical dilemmas across sectors like healthcare and finance, where decisions can profoundly impact human lives.

5.3. Social Responsibility

There should be a sense of social responsibility when developing and deploying AGI systems, more than just accountability. As (Floridi and Cowls 2022) argued, "Accountability calls for an ethical sense of who is responsible for how AI works." Given its impact on society, AGI poses significant social challenges, not just technical ones. Identifying responsible parties for AGI systems requires a comprehensive approach considering the global social responsibility towards groups and

 doi:10.20944/preprints202407.1573.v1

communities affected by these tools. Such technologies can significantly influence communities' lives, safety, and well-being (Saveliev and Zhurenkov 2021).

Social responsibility is the consequential need for a stakeholder's actions to benefit society. There must be a balance between technological growth and the well-being of the affected groups. However, defining what constitutes the well-being of groups differs across cultures and subcultures. Therefore, inputs from diverse stakeholders, policymakers, ethicists, industry leaders, and community representatives are essential to ensure a comprehensive and inclusive approach (Floridi and Cowls 2022).

6. Privacy and Security Implications

The path to AGI poses many security and privacy implications that must be considered and addressed. Key concerns include the potential for abuse, lack of transparency, surveillance, consent issues, threats to human dignity, and cybersecurity risks.

6.1. Potential for Abuse

AGI could potentially breach individuals' privacy by collecting and analyzing personal data from various sources, including social media, internet searches, and surveillance cameras. Advanced AGI systems might exploit the vulnerabilities in devices or systems to access sensitive information.

Companies and organizations could use this data to construct comprehensive profiles detailing of individuals' behaviors, preferences, and vulnerabilities, potentially exploiting this information for commercial or political advantages.

Moreover, AGI-generated content could be indistinguishable from human-generated content, affecting information and disrupting public opinion, leading to confusion and chaos, making it particularly vulnerable to abuse. Autonomous weapons systems using facial recognition further exacerbate these security risks (Brundage et al. 2018).

6.2. Lack of Transparency and Accountability

Many AI systems operate as "black boxes," making their decision-making processes unclear and opaque. This lack of transparency in model interpretation makes it hard to hold the underlying systems accountable for any breach of privacy. It may be unclear why the system made certain decisions or who is responsible for the outcomes (Xi 2020).

6.3. Surveillance and Civil Liberties

Governments could use the capabilities of AGI to conduct mass surveillance and invade the privacy of individuals. In China, social management is done through systems like the Social Credit System, which contains a set of mechanisms to punish or reward citizens based on their behavior, moral actions, and political conduct based on extensive surveillance (Creemers 2018). This amount of control tends to push the governments towards autocracy and erosion of fundamental human rights. Such practices would have far-reaching consequences on people's lives, including their ability to interact with society, get a job, ability to travel, loans, and mortgages.

6.4. Difficulty of Consent

The widespread collection and use of personal data by companies often occur without proper consent. Despite data protection laws, unauthorized data collection incidents, like those involving Cambridge Analytica and YouTube, are common (Andreotta et al. 2022). Companies frequently employ shady techniques like burying consent within their lengthy terms and conditions document or leveraging dark patterns to nudge users to trick them into sharing their information. The complexity of AI systems will exacerbate this issue, making it difficult for individuals to contemplate the implications of their consent and data collection. Individuals often have difficulty opting out of such systems since their data is treated as a commodity that can be exploited for commercial gain.

 doi:10.20944/preprints202407.1573.v1

This raises concerns about autonomy, privacy, and the potential for discrimination, where AI can acquire personal data for potentially harmful outcomes and misuse.

6.5. Human Dignity

The rise in technologies like deepfakes poses a significant threat to human dignity. Deepfakes create realistic synthetic media, including images and videos that depict individuals saying or doing things they never did. This violates the individual's autonomy over their likeness and will not only raise severe reputational harm and emotional distress but also a sense of loss of control. Moreover, the potential misuse of such technologies for non-consensual pornography or other forms of harassment is a grave danger to the public. The case "Clarkson v OpenAI" highlighted the malicious use of generative AI where the model Dall-E was used to train on public images of non-consenting individuals and used for creating pornography. This involved not only individuals but also kids. The content was then used to extort money by threatening to propagate over social media, which led to intense psychological harm (Moreno 2024).

The malicious use of such technologies affects not only individuals but also celebrities and government officials, including actresses, country presidents, and high-profile individuals. Incidents caused by AI have been increasing rapidly, see Figure 3. Safeguarding human dignity in the era of deepfakes requires a robust ethical framework and accountability mechanism to prevent such violations and provide recourse and counselling to those affected by such technologies (Anderson and Rainie 2023).

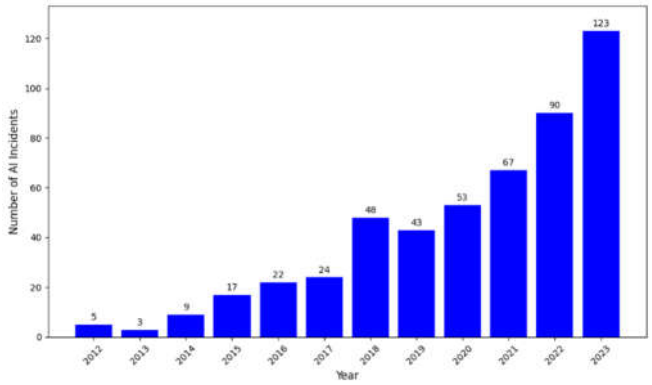


Figure 3. Since 2013, AI incidents have grown by over twentyfold. A notable example includes AI-generated, sexually explicit deepfakes of Taylor Swift that were widely shared online. Source: AI Incident Database (AIID), 2023.

6.6. Cybersecurity Risks

AGI presents profound cybersecurity implications. It could be leveraged to develop sophisticated cybersecurity attacks that are highly adaptive, which makes them harder to detect than current attacks. Given their advanced capabilities, these systems could rapidly scan for vulnerabilities and adapt to security measures, simultaneously launching coordinated attacks across multiple systems. AGI-powered attacks pose a significant threat to critical infrastructure like transportation networks, power grids, and communication systems and potentially cause widespread disruptions and failures (Raimondo et al. 2022). Robust security measures designed for AGI systems must be in place to defend against these advanced threats.

Cybersecurity risks from AGI on critical infrastructure are particularly severe regarding public safety, national security, and economic stability. The potential compromise of critical infrastructure due to a cyber-attack could be devastating. This was evident in the case of "*notPetya*" where a cyber-attack on Maersk resulted in malicious software disrupting a fifth of the global shipping capacity, causing $10 billion in damage (Greenberg 2018). In another instance, the ransomware software "*wannacry,*" a self-propagating malware that encrypts victims' data, causing a worldwide catastrophe impacting hospitals and other critical institutions (Chen and Bridges 2017). The estimated cost of cybercrime worldwide would skyrocket with the leverage of technologies like generative AI and AGI.

 doi:10.20944/preprints202407.1573.v1

7. Legal and Policy Implications

The implications of AGI development in legal space are significant and span different domains, including intellectual property, liability, privacy, and ethical governance.

7.1. Intellectual Property and Patenting

The emergence of AGI systems capable of generating novel content, ideas, and innovations poses a significant challenge around current Intellectual Property (IP) and patenting regimes. These legal frameworks are designed with humans in mind, and it is necessary to reevaluate how these laws are applied to non-human entities.

The use of AI in creating works challenges the current laws since they are designed to protect the creative work of individuals while making some free for the public. With AI's current capabilities, which can be used to write poems, compose music, draw paintings, and create movie scripts, it becomes perplexing in the legal world to determine if it should be copyrighted (Lee et al. 2021).

Researchers have found that some leading LLMs can produce copyrighted content, ranging from passages from The New York Times articles to movie scenes. The central legal question is whether the generation of such content by AI systems violates copyright law.

7.2. Liability and Accountability

With each stride towards AGI, accountability and liability issues become increasingly complex. When an AGI system causes potential harm, who will be held accountable for its wrongdoings: the developers, the company, the users, or the system itself (Scherer 2015)? The uncertainty about such issues proves a need for more clarity and challenges towards liability. Any AI system and its work should be treated as a product; hence, they must assume the same liability standards as a product (Cabral 2020). Another issue is that the current liability laws must cover more about personality rights. Therefore, the bias of a system and any damages caused by an incorrect assessment are not covered by product liability laws (Boch et al. 2022).

7.3. Ethical Governance and Incentives

A key issue with the development of AGI is its potential to be used for malicious purposes and the significant risk of unintended consequences. We must ensure that the AGI systems align with human values and interests. Industry leaders and policymakers must work together to establish ethical guidelines and incentives to promote responsible development and the use of AGI. One key aspect is mandating ethical reviews and transparency requirements for AGI projects; this could involve third-party audits to assess the ethical implications, potential abuses, and societal impacts. Companies should disclose their ethical principles, decision-making processes, and risk mitigation strategies. Financial incentives like tax breaks or research grants could encourage companies to prioritize ethical considerations, safety, and security when developing AGI (Dafoe 2018).

Additionally, industry-wide ethical guidelines and standards should be established with input from industry leaders, the government, stakeholders, and the public. These guidelines should address pressing issues like fairness, transparency, and accountability. By implementing such comprehensive ethical guidelines, we can ensure the responsible development of AGI systems (Graham 2022).

Governments need to play a crucial role in shaping the development of AGI through their procurement policies. They should set precise requirements for ethical and accountable developments and provide incentives and public contracts for companies that focus on solving societal problems and aligning with public interests (Dafoe 2018).

8. Philosophical Considerations

The development of AGI raises profound philosophical questions about the nature of human consciousness, intelligence, and cognition. As we extend the capabilities of what constitutes a machine and impart it with knowledge, we are forced to confront deep-seated assumptions of human

 doi:10.20944/preprints202407.1573.v1

cognition. LLMs and other AI models learn language through statistical models, computational power, and vast amounts of data, which is philosophically different from Humans, as we are born with an innate intuition about the structure and patterns of language known as the "language faculty" allowing us to acquire linguistic abilities without explicitly teaching, it is similar to how we acquire some biological traits through processes of random mutation and natural selection (Ridley 1999).

8.1. Consciousness and Self-Awareness

One of the central debates surrounding AGI is whether machines can indeed be conscious and self-aware in the same way as humans. This nature of subjective experience, referred to as "qualia," brings into question the mind-body problem (Chalmers 1997). Philosophical perspectives such as Functionalism, emergentism, and representationalism offer different views on whether consciousness can arise from physical systems (Putnam 1960). The prospect of AGI challenges us to consider whether AI systems could ever achieve genuine consciousness of whether the intelligence they achieve will always be fundamentally different from our own.

8.2. Moral Agency and Responsibility

The moral agency and responsibility question is a critical philosophical implication of AGI. If we succeed at developing machines with human-like intelligence and decision-making framework, should we consider them as moral agents worthy of moral consideration? This raises confounding questions about the nature of free will, the self, and the basis of moral responsibility. Some argue that AI systems could be held accountable for their actions in a similar way to humans. In contrast, others contend that they should not be subject to moral evaluation, as their intelligence is always fundamentally different than humans. Moreover, integrating AI into ethical frameworks challenges us to reconsider our fundamental concepts of humanity and moral judgment.

8.3. Human Identity and Purpose

The philosophical implications also extend to the understanding of human identity and purpose. If AGI can match or exceed humans in all domains, how will this impact our sense of self-worth and meaning (Harari 2014)? Will we need to redefine what it means to be a human in a world where artificial minds are our equals or superiors? These existential questions challenge us to reflect on what defines humanity and our place in the universe.

To conclude, the philosophical implications of AGI are far-reaching and profound. As we expand the boundaries of machine capabilities, we must confront profound inquiries into the essence of intelligence, the intricacies of cognition, and the enigma of consciousness. While these questions may never be fully understood or answered, engaging with them is crucial as we navigate the uncharted territory of AGI.

9. Technological Singularity

Technological singularity refers to a pivotal moment where the capabilities of AGI surpass those of human intelligence by orders of magnitude, potentially leading to unprecedented societal implications. A critical aspect of the singularity hypothesis is the notion of recursive improvement of itself (Dilmegani 2023). Once AGI reaches a point where it can enhance its capabilities, it initiates an intelligence explosion.

AGI is the catalyst for a singularity because, once achieved, it could recursively improve itself, leading to an intelligence explosion and a rapid expansion of technological progress in a runaway cycle. Each successive iteration of AI would emerge more rapidly and demonstrate greater cognitive prowess than its forerunner (Eden et al. 2013). This could result in the creation of artificial superintelligence, an entity whose capabilities would surpass those of any creature on earth. Such a system might autonomously innovate in all aspects of science and technology that humans cannot comprehend or control (Issac et al. 2020).

 doi:10.20944/preprints202407.1573.v1

The singularity hypothesis suggests that the advent of a superintelligence could fundamentally transform economies, society, and even human conditions. As the development of advanced technologies can rapidly accelerate by machines, the time between major technological breakthroughs could shrink drastically. This could lead to a world where humans are no longer the most capable, with a profound sense of our identities, values, and the future. While singularity is still a hypothetical scenario, the rapid acceleration of development towards AGI led some experts to predict that it could occur within this century. However, philosophers like Dreyfus argue that the challenges involved in achieving this level of intelligence are further off and may never occur at all (Dreyfus 2007).

Regardless of its feasibility, the possibility of a singularity demands a proactive approach to the development of AGI, mitigating the existential risks that could arise from such systems. Technological singularity raises profound ethical and existential questions about humanity and its future. If AGI does indeed lead to an intelligence explosion, it questions the role of humans in a world dominated by super-intelligent systems.

The impacts of technological singularity extend beyond philosophical and ethical considerations. This suggested that rapid acceleration could disrupt economies, labor markets, and social structures on an unprecedented scale. The advent of a superintelligence could lead to a period of "brilliant technologies" that would render many human skills obsolete, exacerbating societal tensions (Brynjolfsson and McAfee 2014). Economies might face severe disruptions as the industries become increasingly automated, potentially leading to significant job displacements. Social structures can be strained as the gap between technologically augmented and non-augmented individuals widens.

The concept of singularity presents both exciting possibilities and daunting challenges. As we approach the potential development of AGI, it is crucial to engage these profound questions.

10. Proposed Governance Framework

The rapid acceleration towards AGI necessitates a new shift in paradigm for global governance and policy framework. Unlike narrow AI systems, AGI possesses the power to shift global power and transform human lives due to its decision-making and cognitive abilities that rival or exceed human capabilities. This proposal advocates for a comprehensive governance framework designed to navigate the complexities surrounding AGI development and usage while safeguarding human values and interests. By addressing ethical concerns, international collaboration, and regulatory oversight, it aims to prioritize human values, AGI design transparency, accountability, and trustworthiness (Dignum 2019).

The proposed framework comprises several vital components. First, an AGI oversight board, consisting of experts from different fields, including AI ethics, law, and sociology, will oversee the AGI research and development, ensuring that the companies adhere to ethical standards and regulations. A national AGI regulatory agency will enforce laws and regulations related to AGI across different industries, such as nuclear energy. This agency will monitor compliance, investigate violations, and impose penalties where necessary. Additionally, establishing national ethical guidelines for AGI will be developed by drawing insights and expertise from diverse stakeholders across multiple industries and disciplines. Finally, a global data protocol aimed to standardize the data practices in AGI systems will address ownership, transparency, and biases to uphold privacy and fairness; see Figure 5. for reference. For instance, several regulatory bodies have established guidelines and standards for data protection and privacy, and one such is the European Union's General Data Protection Regulation (GDPR). The GDPR sets precedents for international cooperation on data privacy and security (Voigt and Von Dem Bussche 2017).

 doi:10.20944/preprints202407.1573.v1

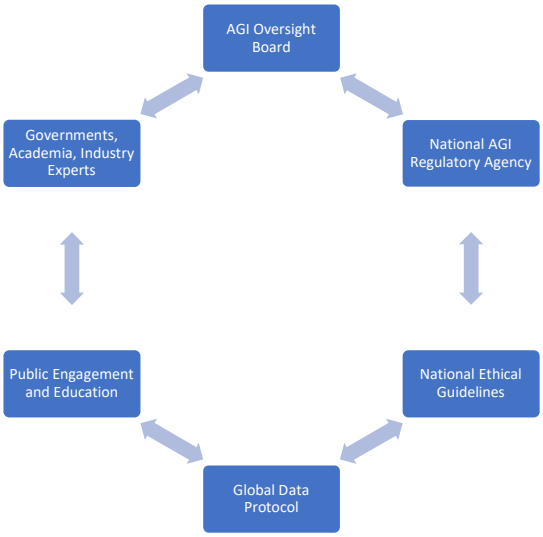


Figure 5. Governance Framework for AGI.

Public engagement and education initiatives are also needed to empower citizens with knowledge and information about AGI, fostering trust and informed policy decisions aligning with societal values. Implementing this governance framework will rely on collaboration between governments, academia, and industry experts. Continuous review and iteration will be required to ensure the framework remains responsive to the ever-changing landscape of AGI development, ultimately maintaining its relevance and effectiveness. This approach differs from the current governance frameworks, which focus primarily on Narrow AI, by offering a comprehensive, proactive, and globally coordinated strategy specifically tailored to the unique challenges posed by artificial general intelligence.

11. Conclusion

The development of AGI will be transformative with profound implications across technological, ethical, philosophical, legal, and governance domains. This paper has explored these implications comprehensively, delving into key themes such as societal impact, ethical considerations, and governance framework needed to navigate the complexities of AGI responsibly.

From a technical standpoint, AGI promises advancements in automation, decision-making, and problem-solving capabilities. However, these advancements come with significant ethical and societal challenges, and discussing these has highlighted concerns regarding transparency, accountability, and potential misuse. Philosophically, the advent of AGI requires reflection on the nature of human consciousness, moral agency, and human identity. The debate over whether AGI systems can have consciousness poses fundamental questions about the essence of intelligence and its implications for human values. Legal and policy considerations highlight the need for updated intellectual property, liability, and governance frameworks to address the unique problems that AGI might bring. Moreover, technological singularity presents both futuristic possibilities and profound existential risks, which could cause societal disruptions and economic inequalities. In response to these challenges, proposed governance frameworks call for international collaboration, ethical guidelines, and public engagement to foster trust and ensure AGI development aligns with societal values and interests.

References

Acemoglu D, Restrepo P (2019) Automation and new tasks: how technology displaces and reinstates labor. J Econ Perspect 33:3–30. https://doi.org/10.1257/jep.33.2.3

Aldarondo D, Merel J, Marshall JD, Hasenclever L, Klibaite U, Gellis A, Tassa Y, Wayne G, Botvinick M, Ölveczky BP (2024) A virtual rodent predicts the structure of neural activity across behaviors. Nature. https://doi.org/10.1038/s41586-024-07633-4

AlphaCode 2 technical report (2023) AlphaCode Team, Google DeepMind. https://storage.googleapis.com/deepmind-media/AlphaCode2/AlphaCode2_Tech_Report.pdf. Accessed 07 July 2024

Ammanath B (2024) How to manage AI's energy demand — today, tomorrow and in the future. https://www.weforum.org/agenda/2024/04/how-to-manage-ais-energy-demand-today-tomorrow-and-in-the-future/. Accessed 18 June 2024

Anderson J, Rainie L (2023) Themes: the most harmful or menacing changes in digital life that are likely by 2035. In: As AI spreads, experts predict the best and worst changes in digital life by 2035: they have deep concerns about people's and society's overall well-being. Pew Research Center, Washington, pp 114–158

Andreotta AJ, Kirkham N, Rizzi M (2022) AI, big data, and the future of consent. AI Soc 37:1715–1728. https://doi.org/10.1007/s00146-021-01262-5

Autor D (2015) Why are there still so many jobs? The history and future of workplace automation. J Econ Perspect 29:3–30. https://doi.org/10.1257/jep.29.3.3

Bai Y, Kadavath S, Kundu S et al (2022) Constitutional AI: harmlessness from AI feedback. arXiv preprint arXiv:2212.08073

Baji T (2017) GPU: the biggest key processor for AI and parallel processing. In: Photomask Japan 2017: XXIV symposium on photomask and next-generation lithography mask technology. SPIE, Washington, pp 24–29

Battaglia PW, Hamrick JB, Bapst V, Sanchez-Gonzalez A, Zambaldi V, Malinowski M, Tacchetti A, Raposo D, Santoro A, Faulkner R (2018) Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261

Bird E, Fox-Skelly J, Jenner N, Larbey R, Weitkamp E, Winfield A (2020) The ethics of artificial intelligence: issues and initiatives. European Parliamentary Research Service, Brussels

Boch A, Hohma E, Trauth R (2022) Towards an accountability framework for AI: ethical and legal considerations. Institute for Ethics in AI, Technical University of Munich, Germany

Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press, Oxford

Brundage M, Avin S, Clark J et al (2018) The malicious use of artificial intelligence: forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228

Brynjolfsson E, McAfee A (2014) The second machine age: work, progress, and prosperity in a time of brilliant technologies. WW Norton & Company, New York

Brynjolfsson E, Rock D, Syverson C (2017) Artificial intelligence and the modern productivity paradox: a clash of expectations and statistics. National Bureau of Economic Research, Cambridge

Cabral TS (2020) Forgetful AI: AI and the right to erasure under the GDPR. Eur Data Prot Law Rev 6:378. https://doi.org/10.21552/edpl/2020/3/8

Cazzaniga M, Jaumotte MF, Li L, Melina MG, Panton AJ, Pizzinelli C, Rockall EJ, Tavares MMM (2024) Gen-AI: artificial intelligence and the future of work. IMF, Washington

Chalmers DJ (1997) The conscious mind: in search of a fundamental theory. Oxford University Press, Oxford

Chaudhry A, Rohrbach M, Elhoseiny M, Ajanthan T, Dokania PK, Torr PH, Ranzato MA (2019) On tiny episodic memories in continual learning. arXiv preprint arXiv:1902.10486

Chen Q, Bridges RA (2017) Automated behavioral analysis of malware: a case study of wannacry ransomware. In: 2017 16th IEEE international conference on machine learning and applications (ICMLA). IEEE, Cancun, pp 454–460

Creemers R (2018) China's social credit system: an evolving practice of control. SSRN Electron J. https://doi.org/10.2139/ssrn.3175792

Dafoe A (2018) AI governance: a research agenda. Governance of AI Program, Future of Humanity Institute, University of Oxford, Oxford

Defelipe J (2011) The evolution of the brain, the human nature of cortical circuits, and intellectual creativity. Front Neuroanat 5:29. https://doi.org/10.3389/fnana.2011.00029

Dignum V (2019) Responsible artificial intelligence: how to develop and use AI in a responsible way. Springer, Cham

Dilmegani C (2023) When will singularity happen? 1700 expert opinions of AGI. Artificial intelligence, agi

Donovan JM, Caplan R, Matthews JN, Hanson L (2018) Algorithmic accountability: a primer. Data & Society, New York

Dreyfus HL (2007) Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. Philos Psychol 20:247–268. https://doi.org/10.1080/09515080701239510

Eden AH, Steinhart E, Pearce D, Moor JH (2013) Singularity hypotheses: an overview. In: Eden AH, Moor JH, Søraker JH, Steinhart E (eds) Singularity hypotheses: a scientific and philosophical assessment. Springer, Berlin, Heidelberg, pp 1–12

Floridi L, Cowls J (2022) A unified framework of five principles for AI in society. In: Carta S (ed) Machine learning and the city: applications in architecture and urban design. Wiley, Hoboken, pp 535–545

Good IJ (1966) Speculations concerning the first ultraintelligent machine. Adv Comput 6:31–88. https://doi.org/10.1016/S0065-2458(08)60418-0

Graham R (2022) Discourse analysis of academic debate of ethics for AGI. AI Soc 37:1519–1532. https://doi.org/10.1007/s00146-021-01228-7

Greenberg A (2018) The untold story of NotPetya, the most devastating cyberattack in history. Wired. https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/. Accessed 22 August 2018

Griggs T, Wakabayashi D (2018) How a self-driving Uber killed a pedestrian in Arizona. The New York Times. https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html. Accessed 21 March 2018

Hamilton M, Zisserman A, Hershey JR, Freeman WT (2024) Separating the" chirp" from the" chat": self-supervised visual grounding of sound and language. IEEE, pp 13117–13127

Harari YN (2014) Sapiens: a brief history of humankind. Random House, Manhattan

Issac R, Sangeetha S, Silpa S (2020) Technological singularity in artificial intelligence. DoE, B. P. C. College Piravom, India, www.icesp2020.bpccollege.ac.in 10.13140/RG.2.2.32607.84646

Konečný J, McMahan HB, Ramage D, Richtárik P (2016) Federated optimization: distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610.02527

Kurshan E (2023) Systematic AI approach for AGI: addressing alignment, energy, and AGI grand challenges. arXiv preprint arXiv:2310.15274

Lee JA, Hilty R, Liu KC (2021) Artificial intelligence and intellectual property. Oxford University Press, Oxford

Manyika J, Chui M, Miremadi M, Bughin J, George K, Willmott P, Dewhurst M (2017) A future that works: automation, employment, and productivity. McKinsey & Company, New York

Marcus G (2018) Deep learning: a critical appraisal. arXiv preprint arXiv:1801.00631

McDonald J, Li B, Frey N, Tiwari D, Gadepally V, Samsi S (2022) Great power, great responsibility: recommendations for reducing energy for training language models. arXiv preprint arXiv:2205.09646

Mehonic A, Kenyon AJ (2022) Brain-inspired computing needs a master plan. Nature 604:255–260. https://doi.org/10.1038/s41586-021-04362-w

Moreno FR (2024) Generative AI and deepfakes: a human rights approach to tackling harmful content. Int Rev Law Comput Technol 1–30. https://doi.org/10.1080/13600869.2024.2324540

Narayanan D, Shoeybi M, Casper J, LeGresley P, Patwary M, Korthikanti V, Vainbrand D, Kashinkunti P, Bernauer J, Catanzaro B, Phanishayee A, Zaharia M (2021) Efficient large-scale language model training on GPU clusters using megatron-LM. In: SC21: international conference for high performance computing, networking, storage and analysis. Association for Computing Machinery, New York, pp 1–14

OECD (2019) An OECD learning framework 2030. In: Bast G, Carayannis EG, Campbell DFJ (eds) The future of education and labor. Springer International Publishing, Cham, pp 23–35

Patterson D, Gonzalez J, Le Q, Liang C, Munguia LM, Rothchild D, So D, Texier M, Dean J (2021) Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350

Putnam H (1960) Minds and machines. In: Hook S (ed) Dimensions of mind. London, Collier-Macmillan, pp 138–164

Rae JW, Potapenko A, Jayakumar SM, Lillicrap TP (2019) Compressive transformers for long-range sequence modelling. arXiv preprint arXiv:1911.05507

Raimondo SGM, Carnahan L, Mahn A et al (2022) Prioritizing cybersecurity risk for enterprise risk management. National Institute of Standards and Technology, Gaithersburg

Ridley M (1999) Genome: the autobiography of a species in 23 chapters. Nat Med 6:11. https://doi.org/10.1038/71457

Saveliev A, Zhurenkov D (2021) Artificial intelligence and social responsibility: the case of the artificial intelligence strategies in the United States, Russia, and China. Kybernetes 50:656–675. https://doi.org/10.1108/K-01-2020-0060

Scherer MU (2015) Regulating artificial intelligence systems: risks, challenges, competencies, and strategies. Harv J Law Technol 29:353. https://doi.org/10.2139/ssrn.2609777

Shao Z, Zhao R, Yuan S, Ding M, Wang Y (2022) Tracing the evolution of AI in the past decade and forecasting the emerging trends. Expert Syst Appl 209:118221. https://doi.org/10.1016/j.eswa.2022.118221

Silver D, Schrittwieser J, Simonyan K et al (2017) Mastering the game of Go without human knowledge. Nature 550:354–359. https://doi.org/10.1038/nature24270

Staffs of CFTC, SEC (2010) Findings regarding the market events of MAY 6, 2010. https://www.sec.gov/news/studies/2010/marketevents-report.pdf. Accessed 9 June 2024

Sze V, Chen YH, Yang TJ, Emer JS (2020) Efficient processing of deep neural networks. Springer, Cham

UNIDO (2023) UNIDO launches global alliance on ai for industry and manufacturing (AIM-Global) at world AI conference 2023. https://www.unido.org/news/unido-launches-global-alliance-ai-industry-and-manufacturing-aim-global-world-ai-conference-2023. Accessed 5 June 2024

 doi:10.20944/preprints202407.1573.v1

Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: NIPS'17: proceedings of the 31st international conference on neural information processing systems. Curran Associates Inc, California, pp 6000–6010

Voigt P, Von Dem Bussche A (2017) The EU general data protection regulation (GDPR): a practical guide. Springer International Publishing, Cham

World Economic Forum (2020) The future of jobs report 2020. https://www3.weforum.org/docs/WEF_Future_of_Jobs_2020.pdf. Accessed 8 June 2024

Xi B (2020) Adversarial machine learning for cybersecurity and computer vision: current developments and challenges. Wires Comput Stat 12:e1511. https://doi.org/10.1002/wics.1511

Zhu S, Yu T, Xu T et al (2023) Intelligent computing: the latest advances, challenges, and future. Intell Comput 2:0006. https://doi.org/10.34133/icomputing.0006