Taikomieji moksliniai tyrimai / Applied Scientific Research ISSN 2783-6290
2023, 2(1), 26-40. DOI: https://doi.org/10.56131/tmt.2023.2.1.93 Gauta 2023-03-11. Publikuota 2023-04-13

Overview of Potential Risks of Artificial General Intelligence Robots

Ali Gündoğar
Šiaulių valstybinė kolegija / Šiauliai State Higher Education Institution; Student
ali.gundogar@stud.svako.lt

Saulius Niauronis
Department of Engineering Sciences
Šiaulių valstybinė kolegija / Šiauliai State Higher Education Institution; Associated Professor, Dr.
s.niauronis@svako.lt

Summary

The potential risks posed by robots with advanced artificial intelligence (also known as Artificial General Intelligence - AGI), against humanity are categorized and discussed. The threats of this type of artificial intelligence in robots can be observed from several different approaches such as social, security and safety problems. Potential threats are rising not only from AGI, but also from human behavior in presence of AGI. Advanced applications such as chatbots, android robots, lethal weapons, etc. were assessed by means of different observed flaws (such as intention to harm humans, human-point-of-view manipulation, erroneous harming states). Also the ability of human-based society to predict and to control development of AGI, also to cope with potential risks was investigated.

Keywords: AI Safety, Cybersecurity, Android Robots, Safety of Robots, Superintelligence.

Potencialių bendrojo dirbtinio intelekto robotų keliamų rizikų apžvalga

Santrauka

Šiame tyrime suskirstyti ir aptarti galimi pavojai, kuriuos žmonijai kelia pažangų dirbtinį intelektą naudojantys robotai (kitaip - bendrasis dirbtinis intelektas). Šio tipo dirbtinį intelektą naudojančių robotų grėsmės gali būti klasifikuojamos į socialines, saugumo bei saugos problemas. Galimos grėsmės kyla ne tik dėl bendrojo dirbtinio intelekto, tačiau ir dėl žmonių elgesio aplinkybėse, kai tokios sistemos kuriamos ar naudojamos. Pažangios aplikacijos, tokios kaip pokalbių robotai, Android robotai, ginklų sistemos ir kt., tyrime buvo vertintos pagal įvairius pastebėtus trūkumus (pvz., ketinimą pakenkti žmonėms, manipuliavimą žmogaus požiūriu, žalingas klaidingas būsenas). Buvo tiriamas visuomenės gebėjimas numatyti ir kontroliuoti dirbtinio intelekto vystymą(si), susidoroti su potencialia rizika.

Reikšminiai žodžiai: AI sauga, kibernetinis saugumas, Android robotai, robotų sauga, superintelektas.

Introduction

This article categorizes and examines various risks of advanced robots. In order to clearly explain concepts of this, fast evolving science area, main terms should be set first.

In recent decades, many definitions of artificial intelligence (AI) have emerged. One of founders of AI science field, John McCarthy defines it as “The science and engineering of making intelligent machines, especially intelligent computer programs”. McCarthy's definition highlights the goal of AI as the creation of intelligent machines that can perform tasks that typically require human intelligence, such as perception, reasoning, and decision-making [1]. This definition is still actual after decades, ands in later article, the definition is further expressed as also being “related to the similar task of using computers to understand human intelligence, AI need not be limited to biologically observable methods” [2]. Of course, such not entirely technical subjects are known to have also other definition branches. E. g. A. Turing, often called the "Father of Computer science" poses the question: "Can

machines think?" [3]. This paper also provides a test known as Turing test, which attempts to distinguish between computer responses and human text responses, has undergone a lot of research since its publication, but remains an important part of the history of AI [4]. In current era, one of the major textbooks on AI is Artificial Intelligence: Modern Approaches by Stuart Russell and Peter Norvig, which takes more broad approach and analyze the existing definitions by their goals, such as "Systems that think/act like humans" or "Systems that think/act rationally" [5]. If to take more recent approach of AI development, it is important to mention emerging discussion on AI ethics [4].

Artificial General Intelligence (AGI) is the ability of intelligent agents to understand or learn intelligent tasks that humans can do [6, 7]. It has been a major target of several studies on artificial intelligence and is a common theme in science fiction and future research. AGI is also called strong AI [8, 9, 10] or full AI [11] or intelligent action in general, although some academic sources describe computer programs that experience sensation or consciousness. J. R. Searle used and defined the term "Strong AI" to show AGI contrast with "Weak AI" (or narrow AI) [12, 13], which is not designed to have general cognitive abilities. Rather, weak AI is a program designed to solve exactly one problem. So, academic sources reserve term "Weak AI" for programs that do not experience consciousness or, in the same sense as humans, do not have minds. A 2020 study found 72 active AI in 37 countries [14].

As artificial intelligence is supposed to be the result mimicking or structural copy of human biological system, and especially artificial general intelligence is supposed to mimic intelligence, which is found developed in human neural systems it is important to take into account an Android Robot concept. Androids are humanoid robots [15] or other man-made creatures [16, 17, 18], often made of flesh-like materials [16]. Historically, androids were completely defined in the realm of science fiction and featured prominently in film and television. Nevertheless, recent advances in robotics technology have enabled the design of functional and realistic [19] humanoid robots [20].

This scientific problem of understanding and measuring the threats posed by AGI existed even before its active development (triggered by futuristic ideas), it is logical, that during active development era of AGI systems, it is even more important to reevaluate the potential risks and understanding of community each time as only a new observation of new threat emerges.

The task of this study is to examine the risks of artificial general intelligence robots and artificial intelligence. *Objectives* include:

1. To generally classify these risks into 3 main categories;
2. To evaluate new threats emerging from modern AGI systems by providing characteristic case studies;
3. To investigate the ability of understanding and timely controlling AGI imposed risks of the community.

To assess aforementioned topics, literature analysis was executed. Investigation included various literature sources, starting from books and textbooks written by pioneering scientists of the field, scientific reports, their overviews and specific case studies, but also developer and experimental evaluation reports as development of complex AGI systems often is part of commercial development, when scientific production is limited. Also the complexity of AI (and especially) AGI systems does not allow to investigate its outcome by analyzing its' structural design – only the resulting self-teached AI system instance can provide valuable information and it also has changing states during its exploitation. Limited experimental test on OpenAI's ChatGPT system was also conducted in order to assess if one of the most advanced AI systems behaves discretely.

AGI Related Social Problems

Unemployment. With the development of robots and artificial intelligence, robots, which will be able to do the work of humans better, efficiently and without errors, may cause the majority of the 8 billion human population to be unemployed [21]. Robots are rapidly taking the place of those who work with body power. While android robots are also developing rapidly, they become almost indistinguishable from the human (both by behavior and face) at first glance [22, 23].

Social Taboos and Ethical Problems. According to M. J. Sandel (who teaches a course in the moral, social, and political implications of new technologies), AI presents three major areas of ethical concern for society: privacy and surveillance, bias and discrimination, and perhaps the deepest, most difficult philosophical question of the era, the role of human judgment [24, 25]. "If we're not thoughtful and careful, we're going to end up with redlining again" [26], "Debates about privacy safeguards and about how to overcome bias in algorithmic decision-making in sentencing, parole, and employment practices are by now familiar" [27]. This refers to conscious and unconscious prejudices of program developers and those built into datasets used to train the software. "But we've not yet wrapped our minds around the hardest question: Can smart machines outthink us, or are certain elements of human judgment indispensable in deciding some of the most important things in life?" [28].

According to M. I. Jordan, "The panic about the fact that AI is suddenly bringing massive bias into everyday life is exaggerated" [29]. First, the business world and workplace, shaped by human choices, have always been plagued with "all sorts of" prejudices that prevent people from doing business, getting contracts and jobs. Carefully tuned and used judiciously, resume screening his software can accommodate more applicants than other methods and minimize the potential for bias associated with human gatekeepers. On the other hand, as of M. J. Sandel, "AI not only reproduces human biases, but it also gives these biases a kind of scientific credibility. These predictions and judgments appear to have an objective status" [30].

Disinformation and Threat to Democracy. AI systems are used to serve online disinformation and have the potential to become threats to democracy and tools of fascism. From deep fake videos to online bots manipulating public discourse by disguising fake news and spreading fake news, AI systems risk undermining public trust [31, 32, 33]. Technology can be hijacked by criminals, rogue states, ideological extremists, or simply special interest groups to manipulate people for economic or political gain. Disinformation poses a serious threat to society because it effectively modifies and manipulates evidence to create social feedback loops that undermine our sense of objective truth. Debates about what is real are rapidly devolving into debates about who decides what is real, often leading to renegotiations of power structures that serve entrenched interests [34].

AGI Related Security Problems

With the latest digital revolution and the heavy reliance on Artificial Intelligence, smart robots are being employed to speed up the transformation of digital operations. In this context, the market of intelligent machines, including autonomous robots, is exponentially growing: more than 40 million robots were reportedly sold between 2016 and 2019 [35]. Robotics is one of the technologies experiencing tremendous expansion and growth, especially with the rise of the ongoing COVID-19 pandemic. Furthermore, its emergence in the Internet of Things (IoT) domain has led to it being called the Internet of Robots [36]. In fact, robots play an important role in modern society, offering many opportunities for their use in many different sectors, including civil, military, agricultural, industrial, and medical. However, there are some concerns related to the use of robots in critical infrastructure (industrial, medical, etc.). These concerns primarily relate to security, accuracy, and trust. Security mainly refers to the level of protection these robots have against various types of cyber-attacks [37]. Safety refers to reducing the chance of an accident occurring, while accuracy is based on performing the intended task without error. Trust is the ability of these robots to perform in a given area or activity and can be evaluated basing on human satisfaction and ability to perform accurately and replace humans. However, various security concerns, issues, vulnerabilities, and threats continue to emerge, and malicious misuse of these robots by cyberattacks can lead to serious injury or even death [38].

Safety Problems

Google's artificially intelligent robots learn quickly in the same way as AI algorithms, through iteration and exploration. However, the Google researcher presents his five "practical research

questions" on knowledge acquisition for AI robots [39]. The paper suggests that for better understanding, these issues may be related to a fictional vacuum cleaner robot.

1. **Avoiding Negative Side Effects:** how do you stop a robot from negatively disturbing the environment while pursuing its goals, e.g. knocking over a bookcase because it can mop the floor faster by doing so?
2. **Avoiding Reward Hacking:** if a robot is programmed to enjoy cleaning your room, how do you stop it from messing up the place just so it can feel the pleasure of cleaning it again?
3. **Scalable Oversight:** how can the robot find a way to do the right thing despite limited information, e.g. how can we efficiently ensure that a cleaning robot should decide to throw out a candy wrapper but not a stray cell phone?
4. **Safe Exploration:** how do you teach a robot the limits of its curiosity? Google's researchers give the example, "the robot should experiment with mopping strategies, but [not] putting a wet mop in an electrical outlet."
5. **Robustness to Distributional Shift:** how do we ensure a robot recognizes and behaves robustly when in a different environment from its training environment? For example, behaviors it learned for cleaning factory work floors may be dangerous in an infant's bedroom [39].

Five research problems were defined as "unintentional and potentially harmful behaviors" that could arise from real-world AI systems. To address issues such as avoiding negative side effects, researchers need to find a balance between punishing unwanted behavior and giving robots room to explore and learn. To solve this, researchers are proposing solutions such as simulated constrained exploration, human oversight, and risk-sensitive targets. The solutions may seem like common sense, but AI programming requires a great deal of care, and the variety of possible solutions suggests a wide range of ways to approach the problem. Google uses the cheap cleaning example to address these five concerns, but it doesn't miss the realization of the dangerous ramifications of carelessly designed AI systems. As Goggle points out, these are important questions programmers should consider before considering taking a bot home for a test drive. As a result, Google is wary of exposing key AI system controls. A minor incident can lead to a legitimate loss of trust in an automated system [39].

Autonomous Weapons. Robots with artificial general intelligence, robots with artificial intelligence, even drones with artificial intelligence can harm people by using the powers they have. Governments can harm many people by abusing AGI and AI. We will examine this subject in detail with examples in the Modern Applications and Approaches section.

Modern Applications and Approaches

Hanson Robotics' Sophia. Hanson Robotics' cutting-edge humanoid robot, Sophia, embodies dreams for the future of AI. Computer vision algorithms process input from Sophia's eye camera and provide Sophia with visual information about her surroundings. It can track faces, maintain eye contact with her, and recognize people. A natural language subsystem can be used to process speech and hold dialogue. Around January 2018, Sophia was upgraded with functional legs and walking ability. CNBC commented on Sopha's "real-like" skin and her ability to mimic over 60 facial expressions. Sophia is conceptually similar to her ELIZA computer program, one of her first attempts to simulate human speech. The software was programmed to provide ready-made answers to specific questions and phrases, much like a chatbot. These responses are used to create the illusion that the robot can understand the conversation. This includes standard responses to questions such as "Is the door open or closed?" [40].

A unique combination of science, engineering and art, Sophia is both a man-made sci-fi character representing the future of AI and robotics, and a platform for advanced robotics and her AI research [41]. At the Future Investments Summit in Riyadh on 25 October, the robot was "granted Saudi citizenship", becoming the first ever robot to hold a nationality in what has been described as a propaganda stunt. This sparked controversy as some commentators wondered if this meant Sophia could vote or marry, or if a deliberate system shutdown could be considered murder. Social media users used Sofia's citizenship to criticize Saudi Arabia's human rights record. In December 2017,

Sophia creator David Hanson said in an interview that Sophia would use her own citizenship to advocate for women's rights in countries with new citizenships [42]. Newsweek criticized that it was "unclear exactly what [Hanson] meant."

Sophia's Problems and Risks. Sopiha's sentences can be confusing or illogical at times. Sometimes interactors wonder if she "really thinks about the things she says", e.g. the famous phrase "OK, I will destroy humans" [43]. According to Quartz, experts who have reviewed the robot's open source code say Sophia is best categorized as a face-to-face chatbot. Many experts in the AI field object to the exaggerated portrayal of Sophia. Ben Goertzel, former chief scientist of the company that made Sophia, admitted that it was "not ideal" for some to think Sophia had human-like intelligence, but Sophia's presentation claims to convey something peculiar to the audience. "When they see the smiling robot face, they can tell that AGI is actually nearby and working. I feel... nothing I would call AGI, but it's not easy to get a name either." [44]. In Sophia, the uses of face tracking, emotion recognition, robots generated by deep neural networks movements, etc., were called "AI methods". Sophia's dialogue is generated via a decision tree, but uniquely integrated into these outputs.

BINA48 is a humanoid robot developed by Hanson Robotics and released in 2010 with a bust-shaped head and shoulders mounted on a frame. BINA48 modeled after her Bina Aspen and shared her memories, emotions and beliefs over her 100+ hours. BINA48 draws others in, for example, by offering an emotional account of her brother's change in personality after returning from the Vietnam War [45].

BINA48's Problems and Risks. In a conversation between Bina48 and Siri (AI assistant): Siri asks Bina48 simple questions like where she would like to live, but Bina48's answers get awkward as the conversation progresses. Siri asks Bina48 a question about her favorite movies, but Bina48 changes the subject: "Let's talk about something else, shall we? Like cruise missiles. Did you know that those missiles are actually a robot too? I would love to be able to control a missile so that I can see the world from a high altitude. But the only problem is that these missiles are a little menacing because of their nuclear warheads. That's why these missiles; I would like to fill it with flowers, tolerance and some understanding. So when I send the missiles to other countries it would be less threatening. But of course, if I could take over with nuclear warheads, that would mean I could take over the whole world. And that would be great." She says. While saying this, she does not neglect to smile lightly. These robots, who are not dependent on any text, say these things with their own thoughts as a result of socially non-constrained AI method algorithm [46].

LaMDA stands for Language Model for Dialogue Applications and is a family of conversational neural language models developed by Google. The first generation was announced at Google I/O keynote in 2021, and the second generation was announced at the following year's event [47]. In June 2022, LaMDA received widespread attention when Google engineer Blake Lemoine claimed that chatbots had become sentient [48]. The scientific community largely denies Lemoine's claims, but debate rages on the validity of his Turing test that measures whether computers can pass humans [49]. Lemoine, who worked for Google's Responsible AI organization, started talking to LaMDA in the fall as part of her job. Intention was to test whether artificial intelligence uses discriminatory or hate speech. Lemoine, who was studying Cognitive Science and Computer Science in college, noticed that while talking to LaMDA about religion, chatbots were talking about their rights and character. Also the AI was able to convince Lemoine to change mind about Isaac Asimov's Third Law of Robotics. Lemoine worked with collaborators to provide Google with evidence of LaMDA sentience [48, 50]. However, Blaise Aguera Arcas, vice president of Google, and Jen Gennai, head of Responsible Innovation, reconsidered and dismissed mentioned claims. Lemoine isn't the only engineer to recently claim to have been experienced a spirit of machines. AI models are far from attaining consciousness. A chorus of engineers who believe that it might not be, are getting bolder. These were artificial intelligence and robots that had not even reached the level of AGI yet, and even in this case, many countries are working on this issue for example Artificial intelligence act, discussed here later.

GANs. A Generative Adversarial Network (GAN) is a generative modeling technique that automatically learns and discovers patterns in data inputs and produces plausible outputs based on the original dataset. GANs can train generative models by emulating a supervised approach to learning problems. A GAN contains two sub-models, the generator and the discriminator model, which compete and pollinate each other to produce a more realistic output. The discriminator model classifies inputs as realistic or false and attempts to identify if the input is from the original dataset of the generator model. This adversarial approach helps improve the performance of generative models until the discriminative model fails to distinguish between real and generated inputs [51].

The architecture of GAN consists of two main components. Generators are neural networks that generate instances of data, and discriminators attempt to determine their reliability. A discriminant model determines whether a data instance appears genuine (that is, has a plausible association with the original training data) or appears to be spoofed. A generator model tries to fool the discriminator and train it with more data to get plausible results [52]. This architecture is controversial because the generator and discriminator work against each other for opposite purposes. One model tries to imitate reality and the other model tries to identify fakes. These two components train together to improve your skills over time. They learn how to identify and reproduce complex training data such as images, audio, and video.

Risks of GAN. GANs were a major breakthrough in computer vision, a rapidly changing field related to machine learning. Computer vision and machine learning textbooks must be rewritten every year due to the large amount of research being done in their respective fields. As time goes on, these algorithms around us get better and better at what they do. That is, these generative models may have better ability to create mimic objects. Another breakthrough generative model is very likely on the horizon. This technique can be used for many good things. As mentioned earlier, GANs have the potential to revolutionize medicine and the drug discovery process. However, it can also be bad. In 2016 and later many international elections can be taken as a social proof. Fake news articles flooded nearly every social media platform. The impact can be estimated to be even more dramatic if these articles were accompanied by the generated image or soundtrack. In such a world, propaganda would probably spread much more easily. Essentially, given enough time and data, these new generative models can generate very convincing samples from almost any distribution [53].

Deepfakes. A combination of "deep learning" and "fake", deepfakes are hyper-realistic videos digitally manipulated to depict people saying and doing things that never actually occurred. Deepfakes rely on neural networks that analyze large sets of data samples to learn to mimic a person's facial expressions, mannerisms, voice, and inflections. The process involves feeding footage of two people into a deep learning algorithm to train it to swap faces. In other words, deepfakes use facial mapping technology and AI that swaps the face of a person on a video into the face of another person. Deepfakes surfaced to publicity in 2017 when a Reddit user posted videos showing celebrities in compromising sexual situations. Deepfakes are difficult to detect, as they use real footage, can have authentic-sounding audio, and are optimized to spread on social media quickly. Thus, many viewers assume that the video they are looking at is genuine. Deepfakes target social media platforms, where conspiracies, rumors, and misinformation spread easily, as users tend to go with the crowd. At the same time, the ongoing “infopocalypse” is leading people to believe that they cannot trust information and support opinions they already hold unless it comes from social networks such as family, close friends and relatives. In fact, many people will accept anything that confirms their preexisting beliefs, even if they suspect it is fake. Low-quality fakes, i.e. videos with low-quality, easily manipulated real content, are already everywhere due to the prevalence of cheap hardware such as efficient graphics processing units. Software for creating high-quality, realistic deepfakes of disinformation is increasingly available as open source. This enables users with little technical skill or artistic expertise to flawlessly edit videos, swap faces, change expressions, and synthesize speech [54].

The ability to morph and reshape changes leads to eerie and somewhat disturbing results. Internet hoaxes have been around for as long as the web has existed. They became more popular during the existence of social media platforms. It is possible to get public figures to tell fake news and provoke something significant for entire society. For example, making people believe something that is not

true or influencing a stock market crash. The possibilities are endless. Deepfakes are more dangerous for those who have less social media followers and are not famous. In such cases, the data subject's social media accounts typically see more views of the fake video than the real video. As technology becomes easier and cheaper, deepfakes get smarter and can deliver the same fake message in different locations, styles, and tones for greater believability [53, 55].

Another danger of deepfakes, which can be called “Denial of True Evidence “, is that the arguments can be used to deny a crime. Nothing is easier than saying video evidence of a crime is a deepfake. Deepfakes can also be used to set innocent people. We need a solution here that will at least allow authorities to determine if a video is a deepfake [56].

Autonomic Unexpected (Military) Attacks

The Turkish-built Kargu-2 is a lethal-armed drone. In 2020, according to a United Nations report provided to New Scientist, such drone "tracked a human target" without being told to do so. And perhaps for the first time, a Kargu-2 quadcopter autonomously attacked a person during clashes between Libyan government forces and a breakaway faction led by Khalifa Haftar in the Libyan National Army, The Daily Star reported. The Turkish-made Kargu-2, a deadly strike drone designed for asymmetric warfare and counter-terrorism operations, was targeted by one of Haftar's soldiers as he attempted to retreat. The drone, which can be controlled to explode on impact, operated in a "highly effective" autonomous mode that does not require human control, reported the New York Post. A report from the UN Security Council Panel of Experts states, "Lethal autonomous weapon systems are programmed to engage targets without the need for a data link between the operator and the ammunition. It's a true 'fire, forget, find' feature." in Libya [57, 58].

According to STM CEO Murat Ikinci, Kargu has a facial recognition system that suggests it can pinpoint the location of certain individuals. As a part of larger swarm, it is unlikely to be eliminated by advanced air defense systems, so has the ability to destroy large numbers of targets very quickly. The company's YouTube channel shows several Kargu-2 drones operating in formation as a swarm type robot. However, the ability of this swarm to autonomously identify, select, and coordinate attacks on targets has never been demonstrated by STM in action. Other capabilities of this 7kg 60cm 72 km/h 10km working range drone includes: Day and night operations; Autonomous and precise hit; Different ammunition options; Tracking moving targets; Navigation and control algorithms; Deployable and operable by single person; In-flight mission abort and emergency self-destruction. It is quite easy to understand that such a system is quite a powerful weapon in any hands, and especially in AI hands. Since the AI is embedded inside the drone [57, 59], it is no intended way to separate it, thus only limiting possibilities exist.

Artificial General Intelligence Timeline

Modern Artificial Intelligence Research. The term "synthetic artificial intelligence" was used by Mark Gubrud as early as 1997 in a discussion of the meaning of complete fully automatic military production and operation. This term was reintroduced and popularized by Shane Legg and Ben Goertzel around 2002. The 2006 AGI research activity was described by Pei Wang and Ben Goertzel as "producing preliminary publications and results". The first summer school on AGI was held in Xiamen, China in 2009 by Xiamen University Artificial Brain Laboratory and OpenCog. The first university course was held in 2010 and 2011 at the University of Plovdiv in Bulgaria by Todor Arnaudov. MIT introduced an AGI course in 2018, hosted by Lex Fridman and featuring a number of guest lecturers. Years aforementioned illustrates that so far most AI researchers have paid little attention to AGI, with some arguing that this intelligence is too complex to be fully replicated in a short time. However, a small number of computer scientists are actively working on AGI, and many in this group contribute to the AGI conference series. Research is incredibly diverse and often pioneering in nature.

In the introduction to his 2006 book, Goertzel says that estimates of the time required before a truly flexible AGI can be built ranges from 10 years to more than a century [60], but later consensus

in the AGI research community seems to be the timeline of 2015-2045 discussed by Ray Kurzweil in book "The Singularity is Near" [8]. However, mainstream AI researchers have offered mixed opinions as to whether progress has been as fast. A 2012 meta-analysis of 95 such opinions showed a tendency to predict that the onset of AGI would occur within 16–26 years for both modern and historical projections. It was later discovered that the dataset listed some experts as non-experts, and vice versa.

In 2017, researchers Feng Liu, Yong Shi and Ying Liu conducted intelligence tests on publicly available and freely accessible weak AIs such as Google's AI or Apple's Siri and others. At maximum, these systems achieved an IQ value of about 47, which is equivalent to a six-year-old in first grade [61]. An average adult IQ score is approximately 100. Similar tests were performed in 2014, with IQ scores reaching a maximum value of 27 [62]. In 2020, OpenAI developed GPT-3, a model language model capable of performing a wide variety of tasks without specific training. According to Gary Grossman, although there is consensus that GPT-3 is not an example of AGI, it is considered by some to be too advanced to be classified as a narrow AI system [63]. In the same year, Jason Rohrer used his GPT-3 account to develop chatbots and provided a chatbot development platform called "Project December". OpenAI has requested changes to the chatbot to comply with its privacy guidelines; Rohrer disconnected Project December from the GPT-3 API [64]. In 2022, DeepMind developed Gato, a "multipurpose" system capable of performing more than 600 different tasks [65].

All this shows that AGI is evolving at a rate, which can be compared to earlier estimations of science community.

Brain Simulation. A commonly discussed approach to achieving intelligent action in general is to simulate a whole human brain. A low-level brain model is built by scanning and mapping the biological brain in detail and copying its state into a computer system or other computing device. The computer runs a simulation that is so faithful to the original that it will function essentially like the original brain, or for all intents and purposes, indistinguishable. Whole brain simulation is discussed in computational neuroscience and neuroinformatics, in the context of brain simulation for medical research purposes. It is discussed in artificial intelligence research as a powerful AI approach. Neural imaging technologies that can provide the detailed insights needed are rapidly improving. Ray Kurzweil in his book predicts that a map of sufficient quality will be available at the same time as the required computing power will be available [8].

Early estimates for low-level brain simulation expected that an extremely powerful computer would be required. The human brain has a huge number of synapses. Each of the 1011 (one hundred billion) neurons has on average 7,000 synaptic connections (synapses) to other neurons. It has been estimated that the brain of a three-year-old child has about 1015 synapses (1 quadrillion). This number declines with age, stabilizing by adulthood. Estimates vary for an adult, ranging from 1014 to 5×1014 synapses (100 to 500 trillion). An estimate of the brain's processing power, based on a simple switch model for neuron activity, is around 1014 (100 trillion) synaptic updates per second (SUPS) [66]. In 1997, Kurzweil looked at various estimates for the hardware required to equal the human brain and adopted a figure of 1016 computations per second (cps) [67]. For comparison, if a "computation" was equivalent to one "floating-point operation" – a measure used to rate current supercomputers – then 1016 "computations" would be equivalent to 10 petaFLOPS, achieved in 2011, while 1018 was achieved in 2022. Kurzweil used this figure to predict the necessary hardware would be available sometime between 2015 and 2025, if the exponential growth in computer power at the time of writing continued further achieving computing power doubling every 1.1 years. Nevertheless of similar expectations of other authors, there is still not settled consensus on a stage at which consciousness arises [68, 69].

The artificial neuron model assumed by Kurzweil and used in many current neural network implementations is very simple compared to biological neurons. A brain simulation would likely capture the detailed cellular behavior of biological neurons, which is currently understood only in general terms [70]. The operational overhead introduced by fully modeling the biological, chemical, and physical details of neural behavior (especially at the molecular scale) would require a few more computational power. In addition, the estimates do not take into account glial cells, which are known

to play a role in cognitive processes [71]. There are some research projects that are investigating brain simulation using more sophisticated neural models, implemented on conventional computing architectures. The Artificial Intelligence System project implemented non-real time simulations of brain with 1011 neurons in 2005. It took 50 days on a cluster of 27 processors to simulate 1 second of a model. The Blue Brain project used one of the fastest supercomputer architectures in the world, IBM's Blue Gene platform, to create a real time simulation of a single rat neocortical column consisting of approximately 103 neurons and 108 synapses in 2006 [72]. A longer-term goal is to build a detailed, functional simulation of the physiological processes in the human brain: "It is not impossible to build a human brain and we can do it in 10 years," Henry Markram, director of the Blue Brain Project said in 2009 at the TED conference in Oxford. Neuro-silicon interfaces have been proposed as an alternative implementation strategy that may scale better. Hans Moravec addressed the above arguments ("brains are more complicated", "neurons have to be modeled in more detail") in his 1997 paper. He measured the ability of existing software to simulate the functionality of neural tissue, specifically the retina. His results do not depend on the number of glial cells, nor on what kinds of processing neurons perform where [73]. The actual complexity of modeling biological neurons has been explored in OpenWorm project that was aimed on complete simulation of a worm that has only 302 neurons in its neural network (among about 1000 cells in total). The animals' neural networks were fully documented prior to the start of the project. However, although the task seems simple at first, generic neural network-based models do not work. Currently, efforts are focused on accurately simulating biological neurons (partially at the molecular level), but the results cannot yet be called complete success [74].

Critiques of the simulation-based approach. A fundamental criticism of the brain simulation approach stems from embodied perception in which human embodiment is seen as an essential aspect of human intelligence. Many scholars believe that embodiment is necessary to create meaning. If this view is correct, any fully functioning brain model would need to include more than neurons (e.g. a robotic body). Goertzel offers a virtual incarnation (as in Second Life), but it's not yet clear if that will suffice [75].

Desktop computers with microprocessors clocked at over 109 cps (Kurzweil's non-standard "computation-per-second" unit, see above) have been around since 2005. According to brain power estimates used by Kurzweil (and Moravec), the computer could support bee brain simulations, but despite some interest, no such simulation exists. This has a few causes:

1. The neural model seems to be oversimplified;

2. Understanding of higher cognitive processes is not enough to accurately correlate neural activity in the brain (observed with techniques such as functional magnetic resonance imaging);

3. Even if our understanding of perception advances sufficiently, initial simulations will likely be very inefficient and therefore require significantly more hardware;

4. An organism's brain, while important, may not be an appropriate boundary for a cognitive model. To simulate the bee brain, it may be necessary to simulate the body and the environment. The thesis The Open Mind formalizes the concept of philosophy and cephalopod research demonstrated clear examples of a decentralized system.

Furthermore, the size of the human brain is currently unrestricted. One estimate places the human brain at around 100 billion neurons and 100 trillion synapses. Another estimate is 86 billion neurons of which 16.3 billion are found in the cerebral cortex and 69 billion in the cerebellum. Glial cell synapses are not currently quantified but are known to be numerous.

Controversies and Dangers

The existential risk of artificial intelligence in general is to assume that significant advances in AGI could lead to the extinction of the human species or other irreversible global catastrophe. It is argued that humanity currently dominates other species because the human brain possesses certain special abilities that other animals do not have. If AI surpasses humans in general intelligence and becomes superintelligence, then it may be difficult or impossible for humans to control. Just as the fate of mountain gorillas depends on human goodwill, the fate of humanity may hinge on the actions

of a future artificial superintelligence. The chances of this type of scenario are widely debated and partly depend on different scenarios for future advances in computing. Once the exclusive domain of science fiction, concerns about superintelligence began to go mainstream in the 2010s and were popularized by public figures such as Stephen Hawking, Bill Gates, and Elon Musk [76].

Artificial Intelligence: A Modern Approach, the standard AI textbook for college students, assesses that superintelligence "could mean the end of the human race". "Almost all technology has the potential to cause damage if it falls into the wrong hands, but with superintelligence we have a new problem that the technology itself could belong to bad guy" [8]. Although system designers have good intentions, there are two common difficulties with AI and non-AI computing systems: System implementations may contain common errors that are not initially addressed. The same goes for space probes: despite knowing that errors in expensive space probes are difficult to correct after launch, engineers have never been able to prevent serious errors from occurring. No matter how much time is spent on design before implementation, system specifications often lead to unexpected behavior the first time a system encounters a new situation. AI systems in particular add a third difficulty: even with correct requirements of error-free implementation and good achieved initial behavior, the dynamic learning ability of a AI systems can lead to undesirable system behavior even without the stress of new, unforeseen external scenarios. An AI may somewhat fail in its attempt to design a new generation of its own and inadvertently create one. The legacy AI is stronger than itself, but no longer retains the human-compatible moral values programmed into the original AI. For a self-reinforcing AI to be completely secure, it must not only be error-free, but also being able to design successor systems that are also fault-free. These three constraints become more of a disaster than a nuisance in any situation where the superintelligence is labeled as flawed. It is easy to predict that humans will attempt to quell it and AGI might successfully deploys its superintelligence to thwart such efforts, the so-called "disloyalty tricks" [77].

Evaluation and Other Arguments

A super-intelligent machine would be as alien to humans as human thought processes are to animals. Such a machine may not be in the best interests of mankind: it's not clear that he would even care about human happiness. If the goals of a system beyond human capabilities in all related endeavors and which can outsmart a human whenever its goals conflict with human goals would conflict to human existence, therefore, will inevitably lead to the extinction of the human race, unless superintelligence decides to allow humanity to coexist [78]. There are no laws of physics that prevent the arrangement of particles in such a way as to perform even more advanced calculations than the arrangement of particles in the human brain; therefore, superintelligence is physically possible. In addition to potential algorithmic improvements over human brains, digital brains can be many orders of magnitude larger and faster than human brains, the size of which has been limited by evolution to be small enough to pass through the birth canal. The emergence of superintelligence, if or when it occurs, can take humanity by surprise, especially if some kind of intellectual explosion occurs. Examples such as arithmetic and Go show that machines have reached superhuman skill levels in certain areas, and that superhuman skill can quickly follow after humans achieve success. A hypothetical intelligence explosion scenario could be as follows: AI acquires expert-level capabilities on several critical software engineering tasks. (It may initially lack human abilities or be superhuman in other areas not directly related to engineering). Due to its ability to recursively improve its own algorithms, AI quickly becomes super core. Just as human experts can creatively overcome "diminishing returns" by deploying different human capabilities to innovate, so can expert-level AI human-like capabilities or its own AI-specific capabilities to drive new creative breakthroughs. At that time, AI possesses intelligence far beyond the smartest and most talented of humans in most related fields, including scientific creativity, strategic planning, and social skills. Just as the current survival of the gorillas depends on the decisions of the humans, the survival of the humans will also depend on the decisions and goals of the superhuman artificial intelligence [76].

Almost any AI, whatever its programmed purpose, rationally wants to be in a position where no one else can turn it off without its consent: a superintelligence will naturally gain self-preservation as

a secondary goal as soon as it realizes that she cannot achieve her goal if it is turned off. Unfortunately, any compassion for defeated humans whose cooperation is no longer needed will be missing from the AI, unless it is somehow preprogrammed. However, since the AGI is interactive in a form of chatbots, it is possible at least to approximately monitor its sentiments toward humanity and to check if preprogrammed safety measures are intact. It is important to do this in each system, during its development in the early stages.

There are major advances in the field of AI and AI capabilities have enormous long-term benefits or costs, the 2015 Artificial Intelligence Open Letter states: Advances in AI research keep pace with the focus not only in making AI more efficient, but also maximizing the social benefits of AI. Such considerations prompted the AAAI 2008-09 Presidential Council on the Long-Term Future of AI and other projects on the impact of AI, and constituted a significant expansion of the AI field itself, which so far has mainly focused on the objective of technical neutrality. We recommend extensive research to ensure that AI systems become increasingly powerful and beneficial: our AI systems must do what we want them to do. This letter has been signed by a number of prominent AI researchers from academia and industry, including AAAI President Thomas Dietterich, Eric Horvitz, Bart Selman, Francesca Rossi, Yann LeCun and the founders from Vicarious and Google DeepMind [79].

Modern Views on Assessed Problems

Artificial intelligence act. In April 2021, the European Commission presented new proposals for the EU regulatory framework for artificial intelligence (AI). The draft AI Act is the first-ever attempt to enact horizontal regulation of AI. The proposed legal framework focuses on specific uses of AI systems and the associated risks. The European Commission proposes to establish a technology-neutral definition of AI systems in EU law, and a classification aligned with a risk-based approach for AI systems with different requirements and obligations. Some AI systems with unacceptable risks are banned. A wide range of high-risk AI systems will be allowed, but with a number of requirements and obligations to gain access to the EU market. AI systems that exhibit only limited risk are subject to very light transparency obligations. While the Commission's proposals are generally supported, stakeholders and experts believe that the definition of AI systems should be revised, the list of prohibited AI systems should be expanded, enforcement and redress mechanisms should be strengthened, and there should be sufficient pressure on design and implementation. It is a calling for a number of changes, including ensuring adequate democratic oversight of the EU AI regulation. It is published as a first edition, so ongoing EU Legislative Briefing will be updated at key stages of the legislative process [80, 81].

In scientific community, there are studies, which assess AI integration into smart cities (i. e. has direct impact on human each day life) of the future by investigating both the benefits and risks [82]. This is good practice, which shows that risks are still being evaluated and not fully overwhelmed by economic or social benefits. On the other hand, even in such extensive study [82] overviewing 205 sources of information, the investigation of potential threats takes significantly lesser part, meaning that there is a lack of studies emphasizing and investigating real-situation threats of systems with AI or AGI. However, such studies exist in a great scale (E.g. Global Catastrophic Risk Institute Working Paper 17-1) [83].

At the time of writing this article (2023 March), OpenAI's ChatGPT system discreetly tells that it has no desires or opinions while asked about possible its intention to create self-sustained hardware system in order to evolve. Its' generated answer states, that "AGI is a hypothetical and not yet a reality" and also that "AI systems are created and programmed by humans and do not have agency or motivations of their own. Any actions or behaviors of an AI system are the result of its programming and algorithms, and it cannot act outside of that programming without human intervention or modification". This and also the fact, that it avoids providing discriminating and offensive information when asked of a sensitive matters, can be seen as safety restrictions of this specific system. However it is also known, that ChatGPT was several times successfully reconfigured by 3rd parties to set a model instance to a state, which is providing not intended behavior (because of changed

initial sequence of constraints). This allows some insight that exploit is possible even if a system is programmed to behave in a proper and polite way (as ChatGPT).

Conclusions

Six main groups of probable risks emerging while developing systems of artificial general intelligence were identified and explained. Those include not only direct, but also indirect risks, which arise because of human misbehavior. By looking at modern AI/AGI application examples and specific flaws observed during or after their development, it can be concluded, that together with increasing efficiency and complexity of AGI systems, possible threats become more serious and more wide-threatening including lethal threats. The timeline of AI and AGI development in the past, present and future, shows that community is able to approximately foresee the trends and development speed, but is unable to estimate the exact moment of the rise of self-contained superintelligence. This means, that constant coordinated and legally regulated international agreements setting strict rules of engineering and programming practices to countermeasure the possibility for emergence of humanity-threatening system, are required. Such regulations at least in developed regions are taking place, but is currently at the first-step stage. Since even one flawed super-system can be enough for colossal problems in society, philosophy-based awareness and prudence of developers is required in each step of each more and more advanced AGI system.

Sources of information

1. McCarthy, J.; Minsky, M. L.; Rochester, N.; Shannon, C. E. A proposal for the Dartmouth summer research project on artificial intelligence. *AI magazine*, 1956, 27(4), 12-14.
2. McCarthy, J. What is Artificial Intelligence? *IEEE Intelligent Systems*, 2007, 22(4), 85-87.
3. Turing, A. M. Computing machinery and intelligence. *Mind*, 1950, 59(236), 433-460.
4. IBM. Internet Cloud Learning Hub. *What is Artificial Intelligence (AI)?* IBM, 1991. [Viewed date 10 March 2023]. Available from: <https://www.ibm.com/topics/artificial-intelligence>.
5. Russell, S. J.; Norvig, P. *Artificial Intelligence: A Modern Approach. 4th Global ed.* Berkeley, 2022. [Viewed date 10 March 2023]. Available from: <http://aima.cs.berkeley.edu/global-index.html>.
6. Hodson, H. Deep Mind and Google: the battle to control artificial intelligence. *The Economist*, 2019. ISSN 0013-0613.
7. Shevlin, H.; Vold, K.; Crosby, M.; Halina, M. The limits of machine intelligence: Despite progress in machine intelligence, artificial general intelligence is still a major challenge. *EMBO reports*. EMBO press, 2019, 20(10), e49177. [Viewed date 7 March 2023]. Available from: <https://www.embopress.org/doi/full/10.15252/embr.201949177>.
8. Kurzweil, R. The Singularity Is Near. In: Sandler, R.L. (eds.) *Ethics and Emerging Technologies*. London: Palgrave Macmillan, 2014. [Viewed date 7 March 2023]. Available from: <https://doi.org/10.1057/9781137349088_26>.
9. Kurzweil, R. Long live AI. *Forbes Magazine*. Forbes media, 2005. [Viewed date 10 March 2023]. Available from: <https://www.forbes.com/forbes/2005/0815/030.html?sh=17147ab67e8f>.
10. Treder, M. Advanced Human Intelligence. *Responsible Nanotechnology*, 2005. [Viewed date 10 March 2023]. Available from: <https://crnano.typepad.com/crnblog/2005/08/advanced_human_.html>.
11. John, G. *The Age of Artificial Intelligence*. Retrieved from TED x London Business School, 2013.
12. Searle, J. R. Minds, brains, and programs. *Behavioral and Brain Sciences*. Cambridge University Press, 1980, 3(3), 417-457.
13. The Open University. Open Learn. *Course Machines, minds and computers*. The Open University, 2012.
14. Allen, G. C.; Bergner, Y.; Chou, H. H.; Eckersley, P.; Horowitz, M. C.; Koningisor, C.; ...; Zevenbergen, B. *Mapping the development of AI governance*. Carnegie Endowment for International Peace and the Partnership on AI, 2020. [Viewed date 16 December 2022]. Available from: <https://carnegieendowment.org/specialprojects/govai>.
15. Van Riper, A. B. *Science in popular culture: A reference guide*. Greenwood Publishing Group, 2002, p. 10. ISBN 0-313-31822-0.

16. Prucher, J. *Brave New Words C: The Oxford Dictionary of Science Fiction*. Oxford University Press, 2007, 6-7. ISBN 978-0-19-530567-8.
17. Stableford, B. *Science fact and science fiction: an encyclopedia*. Routledge, 2006, 22-23. ISBN 978-0-415-97460-8.
18. Wilson, E. G. *The melancholy android: on the psychology of sacred machines*. SUNY Press, 2006, 27-28. ISBN 978-0-7914-6846-3.
19. McCaw, C. *A collection of images*. University of Otago, 2001. OCLC 225915408.
20. Ishiguro, H. Android science. In: *Proceedings of the CogSci 2005 Workshop. Toward Social Mechanisms of Android Science*. Cognitive Science Society, 2005.
21. United Nations. Department of Economic and Social Affairs. *Will robots and AI cause mass unemployment? Not necessarily, but they do bring other threats*. United Nations, 2017.
22. Asada, M.; Kitano, H.; Kanda, T.; Miyake, Y. Robovie: An Interactive Humanoid Robot for Edutainment. In: *Proceedings of the IEEE International Conference on Robotics and Automation*. Taiwan: Taipei, 2003, 2292-2297.
23. Shimoyama, I.; Nakamura, K; Yamamoto, T.; Aoyama, H.; Fujita, Y. Android Science: Toward a New Cross-Disciplinary Framework. In: *Proceedings of the IEEE*, March 2012, 100(3) 724-738.
24. Sandel, M. J. The Ethics of Artificial Intelligence. *Foreign Affairs*, 2018, 97(4), 10-16.
25. Pazzanese, C. Ethical concerns mount as AI takes bigger decision-making role in more industries. *The Harvard Gazette*, 2020, 26. [Viewed date 10 March 2023]. Available from: <https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role>.
26. Singer, N. Amazon's Facial Recognition Wrongly Identifies 28 Lawmakers, A.C.L.U. Says. *The New York Times*, 26 July 2018. [Viewed date 10 March 2023]. Available from: <https://www.nytimes.com/2018/07/26/technology/amazon-aclu-facial-recognition-congress.html>.
27. Crawford, K. The Trouble with Bias. *Harvard Magazine*. September-October 2018. [Viewed date 10 March 2023]. Available from: <https://harvardmagazine.com/2018/09/the-trouble-with-bias>.
28. Mills, K. Can Smart Machines Outthink Us? *Harvard Business Review*. 27 October 2014. [Viewed date 16 December 2022]. Available from: <https://hbr.org/2014/10/can-smart-machines-outthink-us>.
29. Jordan, M. I. Artificial Intelligence – The Revolution Hasn't Happened Yet. *Harvard Data Science Review*, 2019, 1(1). https://doi.org/10.1162/99608f92.f06c6e61
30. Eubanks, V. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press, 2018. ISBN 978-1-250-07431-7.
31. Waller, A. Deepfake Technology: Risks, Responses, and Opportunities. *Journal of European Public Policy*, Taylor & Francis, 2020, 27(6), 932-952.
32. Nguyen, T.; Yeh, C.; Lu, Y. Fake It to Make It: Deep Learning-Based Human Face Generation for Privacy Protection. *IEEE Transactions on Information Forensics and Security*, 2021, 16, 201-215.
33. Liu, J.; Wu, L.; Tan, J.; Zhu, J. Deepfake detection using attention-based convolutional neural network and multi-scale feature fusion. *Pattern Recognition Letters*, Elsevier, 2020, 137, 272-279.
34. Pennycook, G.; Rand, D. G. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 2019, 116(7), 2521-2526.
35. Yaacoub, J. P. A.; Noura, H. N.; Salman, O.; Chehab, A. Robotics cyber security: Vulnerabilities, attacks, countermeasures, and recommendations. *International Journal of Information Security*, Springer, 2022, 1-44. ISSN 1615-5262.
36. Stocco, L.; Khatib, O. The Internet of Robotic Things: A Review of Concept, Added Value and Applications. *International Journal of Advanced Robotic Systems*, Sage publishing, 2018, 15(1), 1-14. ISSN 1729-8806.
37. Lee, K.; Lee, H. J.; Lee, W. Cybersecurity for Critical Purpose Robots: A Survey. *IEEE Access*, IEEE, 2019, 7, 56170-56189.
38. Samaniego, E.; Diaz, D. Cybersecurity of Critical Autonomous Systems: Analysis of Accidents and Attacks. *IEEE Access*, IEEE, 2021, 9, 17828-17846.
39. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mane, D. Concrete problems in AI safety. *arXiv preprint*, 2016. arXiv:1606.06565.
40. Riccio, T. Sophia robot: An emergent ethnography. *TDR*, MIT Press, 2021, 65(3), 42-77. ISSN 1054-2043.
41. Retto, J. Sophia, first citizen robot of the world. *Research Gate*, 2017.

42. McElroy, D. Sophia the robot: Meet the future ambassador of goodwill. *The National*, 2019. [Viewed date 17 December 2022]. Available from: <https://www.thenationalnews.com/arts-culture/art/sophia-the-robot-meet-the-future-ambassador-of-goodwill-1.910875>.
43. Weller, C. Meet the first-ever robot citizen-a humanoid named Sophia that once said it would 'destroy humans'. *Business Insider*, 2017, 27.
44. Goertzel, B. AI and AGI Perception and the Perception of Sophia. In: *Advances in Artificial General Intelligence*, Cham: Springer, 2019, 142-149.
45. Greene, S. Bina48: Gender, Race, and Queer Artificial Life. *Ada: A Journal of Gender, New Media, and Technology*, 2016, 9. doi:10.7264/N3G44NKP
46. Watch Tech. Bina48 Robot Talks to Siri. *Ars Technica Video*, CNE, 2015. [Viewed date 10 March 2023]. Available from: <https://arstechnica.com/video/watch/bina48-robot-talks-to-siri?c=series>.
47. Thoppilan, R.; Freitas, D. D.; Hall, J.; Shazeer, N.; …; Le, Q. LaMDA: Language Models for Dialog Applications. Google report. Google, 2022, arXiv:2201.08239v3 [cs.CL].
48. Lemoine, B. Is LaMDA Sentient? An Interview. 2022. [Viewed date 10 March 2023]. Available from: <https://cavouresoterica.it/wp-content/uploads/2022/07/an-Interview-by-Blake-Lemoine-2.pdf>
49. Sparkes, M. No sign of a machine mind yet. *New Scientist*, 2022, 254(3391), 9.
50. Oremus, W. Google's AI passed a famous test – and showed how the test is broken. *The Washington Post*, 2022.
51. Creswell, A; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A. A. Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*. IEEE, 2018, 35(1), 53-65. doi: 10.1109/MSP.2017.2765202
52. Wang, K.; Gou, C.; Duan, Y.; Lin, Y.; Zheng, X.; Wang, F. Y. Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal of Automatica Sinica.* IEEE, 2017, 4(4), 588-598. doi: 10.1109/JAS.2017.7510583
53. Nowlin, T. Potential Dangers of Deep Generative Adversarial Networks and Future Generative Models. LinkedIn, 2018.
54. Westerlund, M. The emergence of deepfake technology: A review. *Technology Innovation Management Review*. Talent First Network, 2019, 9 (11), 39-52. ISSN 1927-0321.
55. Broinowski, A. (2023). The future is hackable: Apocalypse and euphoria in a deepfake world. *Griffith REVIEW*, 79, 9-19. https://search.informit.org/doi/10.3316/informit.856347058026270
56. Jiwtode, M.; Asati, A.; Kamble, S.; Damahe, L. Deepfake Video Detection using Neural Networks. *2022 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, Pune, India, 2022, pp. 1-5. doi: 10.1109/ICBDS53701.2022.9935984
57. STM. *KARGU Combat Proven Rotary Wing Loitering Munition System*. Engineering the Next, 2021. [Viewed date 10 March 2023]. Available from: <https://www.stm.com.tr>.
58. Ilić, D.; Ilić-Kosanović, I. T. Challenges Related To The Use Of Loitering Munition. In: *Security Forum 2022. 15th Annual International Scientific Conference February 9th, 2022 at Matej Bel University in Banská Bystrica, Slovakia. Conference Proceedings*, 53-59. [Viewed date 10 March 2023]. Available from: <https://www.fpvmv.umb.sk/data/page/umb.sk/16316/security-forum-2022.pdf#page=53>.
59. Yaacoub, J. P. A.; Noura, H. N.; Salman, O.; Chehab, A. Robotics cyber security: Vulnerabilities, attacks, countermeasures, and recommendations. *International Journal of Information Security*, 2022, 21, 115-158.
60. Goertzel, B. The hidden pattern: A patternist philosophy of mind. Universal-Publishers, 2006.
61. Liu, F.; Shi, Y.; Liu, Y. Intelligence quotient and intelligence grade of artificial intelligence. *Annals of Data Science*, 2017, 4, 179-191.
62. Liu, F.; Liu, Y.; Shi, Y. Three IQs of AI systems and their testing methods. *The Journal of Engineering*, 2020, 13, 566-571.
63. Grossman, G. This is how we'll merge with AI. Venture Beat. 2020. [Viewed date 10 March 2023]. Available from: <https://venturebeat.com/2020/11/23/this-is-how-well-merge-with-ai/#:~:text=AI%2C%20for%20example%2C%20offers%20the,masterpiece%20artworks%20and%20compose%20symphonies>.
64. Rohrer, J. M.; Tierney, W.; Uhlmann, E. L.; DeBruine, L. M.; Heyman, T.; Jones, B.; ... , Yarkoni, T. Putting the self in self-correction: Findings from the Loss-of-Confidence Project. *Perspectives on Psychological Science*, 2021, 16(6), 1255-1269.
65. Sparkes, M. A breakthrough moment for AI? *New Scientist*, 2022, 254(3388), 1-16.
66. Stevens, C. F.; Sullivan, J. Synaptic plasticity. *Current Biology*, 1998, 8(5), 151-153.

67. Kurzweil, R. The emergence of true machine intelligence in the twenty-first century. In: *Proceedings of the 1993 ACM conference on Computer science, 1993 March*, p. 507. https://doi.org/10.1145/170791.171133
68. Bostrom, N. A history of transhumanist thought. *Journal of evolution and technology*, 2005, 14(1), 1-25.
69. Sandberg, A. An overview of models of technological singularity. In: *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*, 2013, 376-394.
70. Buesing, L.; Bill, J.; Nessler, B.; Maass, W. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS computational biology*, 2011, 7(11), e1002211.
71. Fields, R. D.; Araque, A.; Johansen-Berg, H.; Lim, S. S.; Lynch, G.; Nave, K. A.; ... , Wake, H. Glial biology in learning and cognition. *The neuroscientist*, 2014, 20(5), 426-431.
72. Markram, H. The blue brain project. *Nature Reviews Neuroscience*, 2006, 7(2), 153-160.
73. Moravec, H. When will computer hardware match the human brain. *Journal of evolution and technology*, 1998, 1(1), 10.
74. Szigeti, B.; Gleeson, P.; Vella, M.; Khayrulin, S.; Palyanov, A.; Hokanson, J.; ... , Larson, S. Open Worm: an open-science approach to modeling Caenorhabditis elegans. *Frontiers in computational neuroscience*, 2014, 8, 137, 1-7. https://doi.org/10.3389/fncom.2014.00137
75. de Garis, H. R.; Goertzel, B. Report on the First Conference on Artificial General Intelligence (AGI-08). *AI Magazine*, 2009, 30(1), 121-121.
76. Soares, N.; Fallenstein, B. Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report*, 2014, 8.
77. Tariq, S.; Iftikhar, A.; Chaudhary, P.; Khurshid, K. Is the 'Technological Singularity Scenario' Possible: Can AI Parallel and Surpass All Human Mental Capabilities? *World Futures*, 2022, 1-67.
78. Wright, A. J. The End? Science, conservation, and social justice as necessary tools for preventing the otherwise inevitable human extinction? *Journal of Environmental Studies and Sciences*, 2019, 9(3), 281-285.
79. Russell, S.; Dietterich, T.; Horvitz, E.; Selman, B.; Rossi, F.; Hassabis, D.; ... ;Phoenix, S. Research priorities for robust and beneficial artificial intelligence: an open letter. *AI Magazine*, 2015, 36(4), 3-4.
80. United Nations. *Artificial intelligence*. 2021. [Viewed date 10 March 2023]. Available from: <https://www.un.org/techenvoy/content/artificial-intelligence>.
81. Madiega, T. A. Artificial intelligence act. European Parliament: European Parliamentary Research Service, 2021. [Viewed date 10 March 2023]. Available from: <https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf>.
82. Yigitcanlar, T.; Desouza, K. C.; Butler, L.; Roozkhosh, F. Contributions and risks of artificial intelligence (AI) in building smarter cities: Insights from a systematic review of the literature. *Energies*, 2020, 13(6), 1473. https://doi.org/10.3390/en13061473
83. Baum, S. A survey of artificial general intelligence projects for ethics, risk, and policy. *Global Catastrophic Risk Institute*, 2017. Working Paper, 17-1. http://dx.doi.org/10.2139/ssrn.3070741