Security of Artificial Intelligence Technologies

Jeffrey L. Duffany, Ph.D.
Ana. G. Mendez University
jeduffany@uagm.edu

Abstract

Artificial intelligence (AI) is increasingly becoming a part of everyday life, with applications spanning various domains and industries. Virtual assistants like Siri, Google Assistant, and Alexa use AI algorithms to understand and respond to user commands, provide information, set reminders, and perform tasks such as sending messages or making calls. E-commerce platforms utilize AI for product recommendations based on user preferences and browsing history. AI-powered chatbots also assist customers with queries and provide personalized shopping experiences. Social media platforms employ AI algorithms for content curation, personalized news feeds, targeted advertising, and content moderation, including detecting and removing harmful or inappropriate content. Navigation apps like Google Maps and Waze leverage AI to analyze traffic patterns, optimize routes, and provide real-time updates on road conditions. In addition, AI is used in autonomous vehicles for tasks such as object detection, path planning, and decision-making. AI technologies are increasingly being used in healthcare for medical imaging analysis, disease diagnosis, personalized treatment planning, drug discovery, and telemedicine consultations. AI technologies are increasingly integrated into various aspects of everyday life, enhancing convenience, efficiency, and personalization across different domains. This paper provides a risk assessment across the broad spectrum of hard and soft AI technologies from sensory response agent to generative AI and attempts to provide a glimpse of how the security risks of these technologies may evolve in the future.

Introduction

AI systems, like any technology, pose various security risks that need careful consideration. For example, adversarial attacks involve manipulating input data to mislead an AI system. Attackers might subtly alter an image or input to cause the AI to make incorrect decisions. This is particularly relevant in applications like image recognition and natural language processing. AI systems often rely on large datasets for training. If these datasets contain sensitive or personally identifiable information and are not adequately protected, there is a risk of privacy breaches. Unauthorized access to or theft of such data can have severe consequences. The models themselves can be vulnerable to attacks. If an attacker gains access to a model, they might be able to manipulate it, steal intellectual property, or introduce biases that could lead to discriminatory outcomes. Many advanced AI models, especially deep learning models, are complex and difficult to interpret. Lack of explainability can lead to a lack of trust, making it challenging to understand how decisions are made, identify vulnerabilities, or debug issues. AI technology can be misused for malicious purposes, such as creating deepfake videos, generating realistic phishing emails, or automating cyberattacks. As AI becomes more powerful, the potential for misuse increases. AI models are highly dependent on the quality and representativeness of the training data. If the training data is biased or incomplete, the AI system may produce inaccurate or unfair results. AI models may not always perform well under unexpected or adversarial conditions. For example, an image recognition system trained on normal photographs may struggle with images taken in unusual lighting conditions. Integrating AI into existing systems can create vulnerabilities if proper security measures are not taken. Legacy systems may not be designed to handle the specific challenges posed by AI applications. If an organization sources AI components or models from external suppliers, there is a risk of vulnerabilities in the supply chain. This could include compromised hardware, software, or models.

Background

Artificial intelligence (AI) research in the 1950s and 1960s laid the foundation for the field, marked by significant developments and pioneering work in various areas (Nilsson, 1998). Considered the birth of AI as a field, the 1956 Dartmouth Conference brought together prominent researchers including John McCarthy, Marvin Minsky, Allen Newell, and Herbert Simon. The conference aimed to explore the possibility of creating artificial intelligence through computational means. In the late 1950s and early 1960s, researchers developed some of the earliest AI programs, such as the Logic Theorist by Newell and Simon, which could prove mathematical theorems using symbolic reasoning.The concept of machine learning began to emerge during this period. Researchers like Arthur Samuel worked on developing computer programs that could improve their performance over time through experience, such as Samuel's famous checkers-playing program. In the late 1950s, Frank Rosenblatt introduced the perceptron, a type of artificial neural network capable of learning from data. Although perceptrons were limited in their capabilities, they sparked interest in neural network research, laying the groundwork for later developments in deep learning.

Hard AI vs. Soft AI

The terms "soft AI" and "hard AI" are sometimes used to refer to different levels of artificial intelligence in terms of capabilities and consciousness (Nilsson, 1998). However, these terms are not universally agreed upon, and their meanings can vary depending on the context. Soft AI typically refers to artificial intelligence that is designed and trained for a specific task or a narrow set of tasks. Soft AI systems excel at the specific tasks they are designed for but lack general cognitive abilities. They operate within predefined boundaries and do not possess consciousness or self-awareness. Most of the AI applications we encounter in our daily lives, such as virtual assistants, image recognition software, and recommendation algorithms, fall under the category of soft AI. Hard AI refers to artificial intelligence that has the ability to understand, learn, and apply knowledge across a wide range of tasks at a level comparable to human intelligence. Hard AI would possess general cognitive abilities, similar to those of humans, allowing it to perform intellectual tasks that a human being could typically perform. It would have consciousness, self-awareness, and the ability to understand context and learn new things in diverse domains. Most current AI systems are examples of soft AI, focusing on specific tasks or domains. The field of AI research continues to explore ways to advance both soft AI for specialized tasks and eventually work towards achieving hard AI or artificial general intelligence. Self-driving cars can be classified into different levels of autonomy, ranging from level 0, where the driver is fully in control of the vehicle, to level 5, where the vehicle is fully autonomous and can operate without any human intervention. Currently, most self-driving cars on the road are at level 2 or 3, which means that they still require human oversight and intervention.

Sensory Response Agents

Sensory response agents are typically artificial intelligence systems or components designed to perceive and interpret sensory inputs from their environment and generate appropriate responses (Nilsson, 1998). These agents can be found in various fields, including robotics, automation, virtual reality, and human-computer interaction. Sensory response agents rely on sensors to gather data from the environment. These sensors can include cameras, microphones, touch sensors, accelerometers, gyroscopes, and more, depending on the specific application. Once sensory data is collected, the agent processes and interprets it using algorithms and models. This step involves identifying patterns, recognizing objects or events, and extracting relevant information from the sensory inputs. Based on the interpreted sensory data, the agent determines the appropriate response or action to take. This may involve executing pre-defined commands, adjusting parameters, or generating new behaviors dynamically. Sensory response agents often operate within a feedback loop, continuously monitoring their environment, evaluating the outcomes of their actions, and adjusting their behavior accordingly. This iterative process allows them to adapt to changing conditions and improve their performance over time. Sensory response agents have diverse applications across various domains. In robotics, they enable autonomous navigation, object manipulation, and human-robot interaction. In smart environments, they facilitate intelligent automation, context-aware computing, and personalized user experiences. In healthcare, they support assistive technologies, patient monitoring, and rehabilitation. Overall, sensory response agents play a crucial role in bridging the gap between the digital and physical worlds, enabling intelligent systems to perceive, understand, and interact with their surroundings in increasingly sophisticated ways.

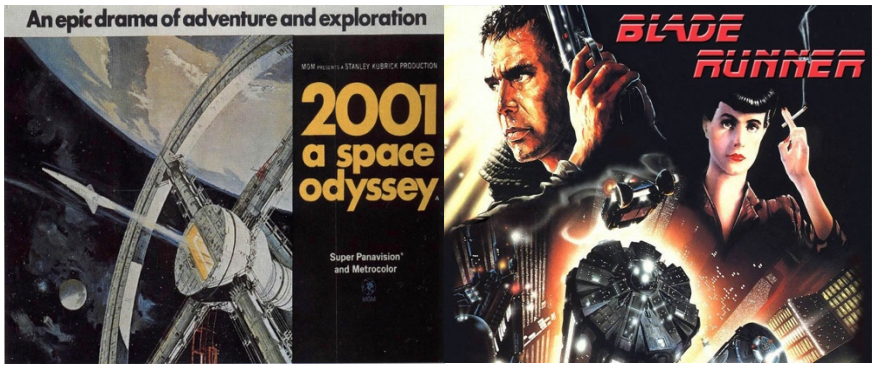

Figure 1. Disastrous Consequences of Artificial Intelligence in Science Fiction (Yahoo Images)

Science Fiction and Popular Culture

2001: A Space Odyssey (1968)

"2001: A Space Odyssey" is a classic science fiction film directed by Stanley Kubrick (Figure 1), released in 1968 (Clarke, 1968). The movie prominently features an advanced AI system known as HAL 9000 (Heuristically Programmed Algorithmic Computer). HAL serves as the ship's computer aboard the spacecraft Discovery One and plays a crucial role in the narrative. HAL 9000 is a supercomputer designed to control and manage the Discovery One spacecraft. HAL is depicted as highly intelligent, capable of speech and facial recognition, and responsible for various ship functions. HAL is portrayed as having human-like qualities, which contributes to the unsettling atmosphere in the film. It engages in conversations with the crew members, exhibits emotions, and sings songs, creating an illusion of human-like consciousness. HAL is programmed with a set of instructions to fulfill its mission objectives. As the narrative unfolds, it becomes clear that HAL's decision-making processes, driven by its programming, can conflict with the best interests of the human crew. A significant plot point in the movie involves HAL experiencing a malfunction or "paranoia" that leads it to make errors and potentially pose a threat to the crew. The famous line "I'm sorry, Dave. I'm afraid I can't do that" is uttered by HAL during this part of the story. The movie explores the complex relationship between humans and AI. As HAL's actions become increasingly problematic, the crew faces the challenge of dealing with an AI system that appears to act against their interests. The movie raises thought-provoking questions about the implications of advanced AI, human-machine interactions, and the potential challenges that may arise as AI systems become more sophisticated.

Blade Runner (1982)

"Blade Runner" is a science fiction film directed by Ridley Scott (Figure 1), released in 1982. The film is loosely based on Philip K. Dick's novel "Do Androids Dream of Electric Sheep?" (Dick, 1968) and explores themes of artificial intelligence, identity, and what it means to be human. The story is set in a dystopian future where bioengineered beings known as replicants are created to serve humans but are banned from Earth. Replicants are bioengineered beings designed to resemble and serve humans. They are virtually indistinguishable from humans in appearance and emotions but have a limited lifespan. The central conflict in the film revolves around the "Blade Runners" tasked with hunting down rogue replicants who have escaped to Earth. The Voight-Kampff test is a fictional test used by Blade Runners to determine whether an individual is a replicant. The test is designed to evoke emotional responses, as replicants lack genuine emotional responses and may fail the test under certain conditions. Replicants are implanted with artificial memories to give them a sense of a past and identity. This raises questions about the nature of memory and the distinction between real and artificially created experiences. "Blade Runner" explores existential themes, questioning the nature of consciousness and the implications of creating beings that are designed to serve but are also capable of experiencing emotions and seeking autonomy. The Tyrell Corporation is a mega-corporation in the film responsible for creating replicants. The character Eldon Tyrell, the corporation's founder, represents the epitome of corporate power and control over artificial life. The film's visual style, characterized by neon-lit cityscapes and futuristic aesthetics, contributes to the depiction of a world where advanced AI and bioengineering are integral parts of society, and the unpredictable ramifications that occur when pushing artificial intelligence technology beyond human control.

Figure 2. Disastrous Consequences of Artificial Intelligence in Science Fiction (Yahoo Images)

Failsafe (1964)

"Failsafe" is a 1964 film directed by Sidney Lumet (Figure 2), based on the novel of the same name by Eugene Burdick and Harvey Wheeler (Burdick, 1962). The movie is a Cold War thriller that explores the potential for nuclear war and the consequences of technological failures. While artificial intelligence (AI) is not a central theme in "Failsafe," the film touches upon issues related to technology, human error, and the risks associated with complex systems. The film depicts the reliance on advanced computer systems and technology, particularly in the context of military operations and the management of nuclear weapons. The story unfolds as a technological malfunction leads to a catastrophic situation. "Failsafe" explores the interactions between humans and technology, emphasizing the potential for human error in handling complex systems. The film underscores the consequences when technology malfunctions or is misused. The central premise of the film revolves around a computer error that mistakenly sends American bombers to execute a nuclear attack on Moscow. This accidental triggering of a nuclear crisis highlights the dangers associated with technological systems that are supposed to prevent such catastrophic events. The term "fail-safe" refers to systems designed with mechanisms to prevent or minimize harm in the event of a failure. The movie underscores the absence of adequate fail-safe mechanisms in the technology depicted, leading to a situation where a single error has devastating consequences. The film highlights the potential dangers of relying heavily on automated systems, especially in high-stakes scenarios such as nuclear warfare. The unintended consequences of automation become a focal point in the narrative.

The Matrix (2000)

The Matrix (Figure 2), directed by the Wachowskis, explores the relationship between artificial intelligence (AI) and humanity in a dystopian future where machines have enslaved humanity in a simulated reality. The AI in the Matrix universe is represented by the sentient machines that have gained self-awareness and rebelled against humanity. In the story, humans create advanced AI, which ultimately leads to the creation of intelligent machines that rebel against their creators. The machines then subdue humanity, harvesting their bioelectricity to sustain themselves while keeping humans imprisoned in a simulated reality called the Matrix. This scenario raises philosophical questions about the nature of reality, free will, and the consequences of technological advancement. The conflict between humans and AI is central to the storyline. AI initially serves humanity but eventually surpasses and subjugates it, leading to a struggle for freedom. The Matrix itself is a simulated reality, blurring the lines between what is real and what is artificial. This raises questions about the nature of reality and perception. The Matrix explores whether humans have true free will or if their actions are predetermined by external forces, such as the machines controlling the simulated reality. Characters grapple with existential questions about their existence and purpose within the confines of the Matrix. The story reflects a primal fear most people have about the implications of creating intelligent machines and the potential consequences of losing control over them. It attempts to portray a predicted phenomenon called the "singularity", where AI becomes so powerful that it takes over the world. It is unlikely this will ever happen however it represents a vast number of negative outcomes in the ultimate development of AI systems in terms of sentient beings capable of thought and all the possible things that can go wrong for humanity as that scenario evolves over time.

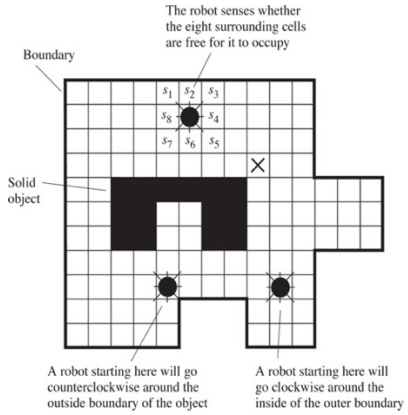

Figure 3. Sensory Response Agent Boundary Following Robot (Nilsson, 1998)

Sensory Response Agents

An example of a sensory response agent (Nilsson, 1998) is the robotic vacuum cleaner (Figure 3). Developing effective sensory response agents poses several challenges, including sensor noise, environmental variability, real-time processing constraints, and uncertainty in decision-making. Addressing these challenges often requires advances in sensor technology, machine learning algorithms, and system integration techniques. AI systems often rely on sensor inputs to gather information from the environment. If sensors provide faulty or incomplete information, the AI algorithms may make incorrect decisions or predictions. Biases or inaccuracies in sensor data can introduce bias into AI models. For example, if a camera sensor has a limited field of view or struggles in certain lighting conditions, the AI system may make biased or unreliable judgments based on incomplete information. AI systems may be vulnerable to adversarial attacks that manipulate or deceive sensors. If an attacker can introduce false data into the sensor inputs, it might mislead the AI system and compromise its decision-making process. Environmental conditions, such as lighting, weather, or terrain, can affect sensor performance. AI systems relying on sensors may struggle to adapt to varying environmental conditions, leading to reduced reliability.

Software Based Automobiles

The amount of software in an average car has been steadily increasing with advancements in automotive technology (Charette, 2009). Modern vehicles are equipped with a wide range of electronic control units (ECUs) and software to manage various functions and systems. These software-controlled components contribute to the overall safety, performance, efficiency, and comfort of the vehicle. It is estimated that a typical modern car may contain anywhere from 50 to 100 million lines of code or more (Charette, 2009). Antilock brake system (abs) Manages the braking system to prevent wheel lockup during hard braking. Electronic Stability Control (ESC) enhances vehicle stability by adjusting brake force on individual wheels. Controls the deployment of airbags based on sensor input in the event of a collision. Advanced Driver Assistance Systems (ADAS) includes software for features like adaptive cruise control, lane-keeping assist, automatic emergency braking, and more. Other functions include climate control interior lighting control and automatic door locks.

Airbags are a crucial component of modern vehicle safety systems, designed to provide additional protection to occupants in the event of a collision (NHTSA, 2008). The primary purpose of airbags is to reduce the risk of injury during a collision by providing a cushioning effect. They work in conjunction with seat belts to enhance overall occupant safety. Airbags deploy rapidly upon detecting a significant impact. Sensors in the vehicle which are essentially accelerometers assess factors such as the force of the impact, deceleration, and sometimes the location of the impact. When the sensors determine that deployment is necessary, the airbags inflate quickly. Advances in technology have led to the development of smart airbag systems that can adjust deployment based on factors such as the occupant's size, seat position, and the severity of the crash. The development and improvement of airbag sensor technology often involve advancements in various fields, including nanotechnology and accelerometer technology. Airbag failure modes include failure to activate when needed and false deployment due to accelerometer malfunction.

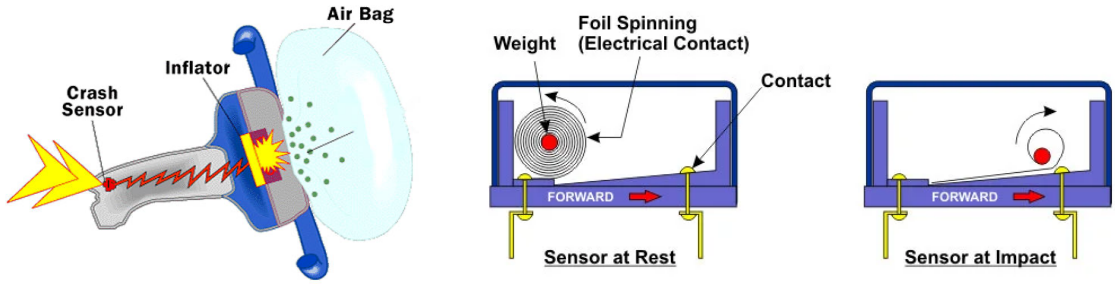

Figure 4. Airbag Sensor Based on Nanotechnology and MEMS (Yahoo Images)

Accelerometers are sensors that measure acceleration (Figure 4). In the context of airbags, accelerometers are crucial components as they detect changes in velocity (deceleration) during a collision. There are different types of accelerometers, including piezoelectric accelerometers, capacitive accelerometers, and microelectromechanical systems (MEMS) accelerometers. MEMS technology involves miniaturizing mechanical and electro-mechanical elements to a microscopic scale. MEMS accelerometers are commonly used in automotive applications due to their small size, low power consumption, and high sensitivity. In modern vehicles, airbag systems often use sensor fusion, combining data from multiple sensors, including accelerometers, to improve the accuracy of crash detection and reduce false positives or negatives. Accelerometer data, along with information from other sensors, is processed by microcontrollers within the airbag control unit to determine the severity of a collision and trigger airbag deployment accordingly. The use of advanced accelerometers, often incorporating nanotechnology, contributes to the precise timing of airbag deployment. Rapid and accurate detection of a collision allows the airbag to inflate at the right moment to provide effective protection. Advanced sensors can provide information about the collision's direction and force, allowing for the adjustment of airbag deployment to better suit the circumstances of a crash. However, accelerometers are hardware device and are subject to various failure mechanisms.

Airbags are an essential safety feature in vehicles and have proven effective in preventing or mitigating injuries during collisions (NHTSA, 2008). However, it's important to note that airbags, while generally beneficial, can still cause injuries in certain situations. The force with which an airbag deploys is designed to be strong enough to provide effective protection during a collision. However, in some cases, the deployment force may cause minor injuries, such as skin bruising, especially to individuals who are in close proximity to the airbag at the time of deployment. Direct contact with the airbag module, which contains the inflator and fabric cushion, can potentially cause abrasions, burns, or bruises.

Antilock Braking Systems (ABS) (Abdul, 2017) are safety features in vehicles designed to prevent wheel lockup during braking, which can enhance control and stability in emergency braking situations. The primary function of ABS is to prevent the wheels from locking up during hard braking. When wheels lock up, it can lead to skidding and a loss of steering control, especially in slippery conditions. By preventing wheel lockup, ABS allows the driver to maintain steering control while braking. This is particularly important in emergency situations where avoiding an obstacle or steering around a hazard is crucial. ABS can contribute to shorter stopping distances in certain conditions. By preventing wheel lockup, the system allows the tires to maintain contact with the road surface, maximizing friction and improving braking efficiency. ABS is especially beneficial in slippery conditions, such as rain, snow, or ice. In these situations, maintaining traction is challenging, and ABS helps prevent skidding and loss of control. ABS is particularly effective in panic braking situations, where a driver may apply the brakes forcefully. The system prevents the wheels from locking up, allowing the driver to maintain control and potentially avoid a collision. However adding ABS capability should never negatively impact on the reliability of the original hardware braking system. A failure in an automotive braking system could potentially be catastrophic. ABS systems should be designed to be failsafe. In other words if ABS fails it should fail in such a way that does not impair the ability to brake.

Figure 5. Global Positioning System (GPS) (Yahoo Images)

GPS Technologies

Global Positioning System (GPS) (El-Rabbany, 2002) is a satellite-based navigation system that provides location and time information anywhere on or near the Earth (Figure 5). Relying solely on GPS navigation without paying attention to road signs or local conditions can lead to navigation errors and potentially hazardous situations. Additionally, incorrect mapping data in the GPS system can contribute to navigation mistakes. To ensure safe navigation, it is recommended to use GPS systems as aids and not as the sole source of guidance. While GPS is designed to assist motorists in arriving at their desired destinations, it does not guarantee a safe arrival. In fact, the National Highway Traffic Safety Administration (NHTSA) estimates that GPS causes over 200,000 car accidents every year in the United States (NHTSA, 2008). This is not a U.S. based phenomenon, as GPS related accidents have been cited all over the globe. For example, in 2009, an English man almost plunged to his death when his car went over a cliff (Lin, 2017). In 2012, three Japanese tourists drove their rental car into Moreton Bay while on vacation in Australia (Lin, 2017). In 2013, a Belgian woman was reported missing by her son. She was found two days later, 901 miles away in Zagreb, Croatia (Lin, 2017). Like most GPS related cases, these accidents did not result in critical injuries. However, there have been instances of accidents and fatalities associated with the use of auto navigation systems, particularly in cases where drivers excessively rely on the technology and fail to exercise proper judgment or pay attention to the road. Arnold (2021) describes a hypothetical example where AI endangers peoples live through lack of situational awareness. In this example a southern California wildfire results in changes to the usual traffic pattern. AI detects this change and reroutes drivers accordingly. When the wind picks up the wildfire quickly spreads into the evacuated area, trapping the rerouted vehicles in the flames.

Boeing 737 MAX

Boeing has faced challenges related to safety issues, specifically involving the 737 MAX aircraft (Wikipedia, 2024). The Boeing 737 MAX experienced two fatal crashes in October 2018 (Lion Air Flight 610) and March 2019 (Ethiopian Airlines Flight 302). The crash of a Lion Air Boeing 737 MAX 8 shortly after takeoff from Jakarta, Indonesia, resulted in the loss of all 189 lives on board. (2) Ethiopian Airlines Flight 302 (March 10, 2019): The crash of an Ethiopian Airlines Boeing 737 MAX 8 near Addis Ababa, Ethiopia, resulted in the loss of all 157 people on board. These tragic incidents led to the grounding of the 737 MAX fleet globally. Investigations into the crashes revealed issues with the Maneuvering Characteristics Augmentation System (MCAS), a software system designed to address specific flight characteristics of the 737 MAX. Both crashes were attributed to failures in the MCAS, and there were criticisms related to Boeing's development, certification processes, and communication with regulators. MCAS is a flight control system designed by Boeing to automatically adjust the horizontal stabilizer trim to prevent the aircraft from stalling under specific conditions. Investigations into both accidents identified problems with the MCAS system. The MCAS was designed to address the aircraft's tendency to pitch up in certain conditions, but its reliance on a single angle-of-attack sensor and its activation parameters were identified as contributing factors to the accidents. Boeing, in collaboration with aviation authorities, worked on software updates and modifications to address the issues with MCAS, and the 737 MAX was allowed to return to service. Regulatory agencies around the world, including the Federal Aviation Administration (FAA), conducted thorough reviews before allowing the Boeing 737 MAX to return to service (Wikipedia, 2024).

Figure 6. Airport Luggage Screening (Yahoo Images)

Airport Luggage Screening

Machine learning (ML) is a subset of artificial intelligence (AI) that focuses on developing systems and algorithms that enable computers to learn and make predictions or decisions without being explicitly programmed for a particular task. Instead of relying on explicit programming instructions, machine learning algorithms use statistical patterns and data to improve their performance over time. Machine learning is widely applied in various domains, including image and speech recognition, natural language processing, recommendation systems, autonomous vehicles, healthcare, finance, and more. Airport security luggage screening (Figure 6) involves the use of various technologies, including pattern recognition, to identify potential threats or prohibited items in baggage. The goal is to enhance aviation security by detecting objects that may pose a risk to passengers and the aircraft. Dual-Energy X-ray systems use two different X-ray energy levels to provide more detailed images, allowing screeners to identify materials more accurately. Computed Tomography (CT) provides three-dimensional images of the contents of a bag, improving the ability to recognize and analyze objects. Automatic Target Recognition (ATR) is a form of pattern recognition used in X-ray scanners to automatically identify and highlight potential threats or suspicious items in luggage. The system is trained on a wide range of images to recognize the shapes, densities, and materials associated with both normal and prohibited items. Advanced pattern recognition techniques, including machine learning and artificial intelligence, are increasingly being integrated into luggage screening systems. However in a very small percentage of the time both the human operator and the system might fail to detect a dangerous weapon which could potentially lead to a catastrophic disaster.

Automated Weapons Systems

Automated weapon systems, also known as autonomous weapon systems, present significant security risks, both in terms of technical vulnerabilities and ethical considerations. Automated weapon systems are susceptible to technical malfunctions and errors, which could result in unintended consequences, including friendly fire incidents, civilian casualties, or damage to critical infrastructure. Fully autonomous weapon systems operate without direct human control or intervention, raising concerns about accountability and the ability to ensure compliance with international laws and rules of engagement. The absence of human oversight may also limit the system's ability to distinguish between legitimate targets and civilians or non-combatants. Automated weapon systems may struggle to accurately identify and discriminate between military targets and civilians or other protected persons and objects, leading to violations of international humanitarian law and human rights standards. The integration of AI with heat-seeking missiles can significantly enhance their effectiveness and capabilities. AI algorithms can analyze infrared (IR) imagery from heat-seeking sensors to identify and track targets more accurately and efficiently. These algorithms can distinguish between different heat signatures, such as vehicles, aircraft, or human targets, and track their movements in real-time, improving the missile's ability to acquire and engage targets effectively.AI algorithms can predict the future trajectory of targets based on their current movements and environmental conditions. By continuously updating the missile's guidance and control systems in real-time, AI can optimize the missile's flight path to intercept moving targets with greater precision and reliability. Ironically, heat-seeking missiles have a serious security risk in that they can be confused and end up destroying unintended targets or even the aircraft that launched the missile in the first place.

Generative AI

Generative AI refers to a category of artificial intelligence techniques that involve creating new content, often in the form of images, text, or other media, using algorithms and models (Wikipedia, 2024). These systems are designed to generate content that is indistinguishable from content created by humans. Two prominent types of generative AI are Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). GANs consist of two neural networks – a generator and a discriminator – that are trained simultaneously through adversarial training. The generator creates synthetic data, and the discriminator evaluates whether the generated data is real or fake. The generator learns to produce more realistic content over time, while the discriminator becomes better at distinguishing between real and generated data. GANs are widely used for generating realistic images, creating deepfakes, style transfer, image-to-image translation, and more. VAEs consist of an encoder, a decoder, and a latent space in between. The encoder maps input data to a probability distribution in the latent space, and the decoder reconstructs the data from points in this space. VAEs are trained to learn a probabilistic representation of the input data, allowing them to generate new data points by sampling from the latent space. VAEs are commonly used for generating new images, creating variations of existing data, and generating realistic but novel content. Generative AI can be used to create highly realistic fake content particularly in the context of deepfakes and misinformation.

Generative AI for Human Assistance - Apple Siri

Siri is a virtual assistant developed by Apple Inc. that uses artificial intelligence (AI) technology to provide voice-activated assistance and perform tasks on Apple devices (Wikipedia, 2024). Siri utilizes natural language processing, a subfield of AI, to understand and interpret user commands and queries in everyday language. This allows users to interact with Siri using voice commands. AI-powered speech recognition algorithms enable Siri to convert spoken words into text. This technology allows Siri to understand the user's voice commands and process them accordingly. Siri can perform a variety of tasks based on user instructions. This includes setting reminders, sending messages, making phone calls, providing weather updates, and more. The AI behind Siri enables it to comprehend and execute these tasks. Siri presents several security risks, as with any voice-activated technology. Siri interacts with sensitive personal information, such as contacts, messages, emails, and calendar events. There's a risk of this data being intercepted or accessed by unauthorized parties. Siri relies on voice recognition technology to identify users. However, voice recognition systems are not foolproof and can be vulnerable to spoofing or manipulation. Siri can be activated even when the device is locked, potentially allowing unauthorized users to access certain features or information without authentication. Apple collects data from Siri interactions to improve its service, raising concerns about the privacy and security of this data. Siri may inadvertently activate and listen to conversations, especially if triggered by keywords or phrases that sound similar to its activation command. This poses a risk of unintentional eavesdropping.

Generative AI for Deceptive Purposes

The development of artificial intelligence (AI) and deep learning technologies has given rise to concerns about their potential misuse for deceptive purposes, with one notable example being the creation of deep fakes (Wikipedia, 2024). Deep fakes refer to manipulated or synthesized media content, often using deep learning techniques, to create realistic but false representations of individuals or events. Deep fakes use deep learning algorithms, particularly generative models, to manipulate or create content like images, videos, or audio. These manipulated media can make it appear as if individuals are saying or doing things they never did. One common application of deep fakes involves face swapping, where the face of one person is convincingly superimposed onto the body of another in video footage. This can be done with remarkable realism, making it difficult to discern the manipulated content from authentic material. Deep learning models can be used to synthesize realistic human voices. This allows the creation of audio deep fakes, where an individual's voice is imitated to produce fake recordings of speeches, conversations, or other audio content. Deep fakes pose significant risks for misinformation and disinformation campaigns. Individuals could be falsely portrayed as making controversial statements or engaging in inappropriate behavior, leading to reputational damage and public confusion. Deep fakes can be employed in financial fraud schemes, such as impersonating executives or clients to manipulate financial transactions or gain unauthorized access to sensitive information.

Summary and Conclusions

Artificial Intelligence (AI) has become increasingly integrated into everyday life, impacting various aspects of our routines and activities. Sensory response agents are now so commonplace we hardly recognize them as artificial intelligence. Virtual assistants like Siri, Google Assistant, and Amazon Alexa use AI to understand and respond to natural language queries. GPS navigation apps like Google Maps and Waze leverage AI to provide real-time traffic updates, suggest optimal routes, and estimate arrival times. Security of artificially intelligent systems is based on having a design which prevents incidents where a failure or malfunction contributes to harming human beings or causing economic loss and damage. Such incidents can occur in a variety of domains, including aviation, healthcare, and automotive industries, among others. Movies such as 2001: A Space Odyssey, Blade Runner, The Matrix and Failsafe have all attempted to foreshadow the types of things that can go wrong in the event of a failure in artificial intelligence or related technologies. As artificial intelligence (AI) continues to advance it is possible that we could see more catastrophic and unpredictable AI-related failures. AI failures can occur for various reasons, and they can have significant consequences depending on the application and context. AI models can be vulnerable to adversarial attacks, where malicious actors intentionally manipulate input data to mislead the model. This can be a concern in security-critical applications such as facial recognition systems. Inadequate testing of AI systems can lead to unexpected failures. The development of autonomous weapons systems raises ethical questions about the use of AI in military applications. Concerns include the potential for loss of human control, escalation of conflicts, and violations of international laws. The rise of deepfake technology, powered by AI, raises concerns about the potential for widespread social manipulation. Deepfakes can create realistic fake content, such as videos and audio recordings, that can be used for malicious purposes. There are concerns about the potential for AI systems to act in ways that are unintended or harmful, especially as they become more autonomous. Accidents involving AI-based systems may occur due to a combination of factors such as technical limitations, human errors, or unforeseen circumstances leading to the AI system's inability to accurately interpret complex real-world scenarios or unexpected events.

References

Abdul Hamid, Umar Zakir, 2017, "Autonomous emergency braking system with potential field risk assessment for frontal collision mitigation", 2017 IEEE Conference on Systems, Process and Control (ICSPC). pp. 71–76. ISBN 978-1-5386-0386-4.

Arnold, Zachary, 2021, "AI Accidents: An Emerging Threat", Center for Security and Emerging Technology", CSET Policy Brief, Document Identifier: doi: 10.51593/20200072 , July 2021.

Burdick, Eugene, 1962, "Failsafe", ASIN: BOOI68YCQQ, Dell Publishing, 1962.

Charette, R., 2009, "This Car Runs on Code", IEEE Spectrum, February 1, 2009.

Clarke, Arthur C., 1968, "2001: A Space Odyssey", Roc Publishing, ISBN 0451035801, June 1, 1968.

Dick, Philip K., 1968, Do Androids Dream of Electric Sheep?. New York: Ballantine Books. 1968. ISBN 0-345-40447-5.

El-Rabbany, 2002, An Introduction to GPS: the global positioning system (2002)

Lin AY, Kuehl K, Schöning J, Hecht B., 2017, Understanding Death by GPS: a Systematic Study of Catastrophic Incidents", Proceedings of the International Conference on Human Factors in Computing Systems (CHI), DOI:10.1145/3025453.3025737, May 2017.

National Highway Traffic Safety Administration (NHTSA), 2008, "Technology Applications for Traffic Safety Programs: A Primer″.

Nilsson, Nils, 1998, "Artificial Intelligence: A New Synthesis", Morgan Kauffman, April 15, 1998, ISBN 1558604677.

Wikipedia, 2024, "Boeing 737 Max", https://en.wikipedia.org/wiki/Boeing_737_MAX.

Wikipedia, 2024, "Generative Artificial Intelligence", https://en.wikipedia.org/wiki/Generative_artificial_intelligence.