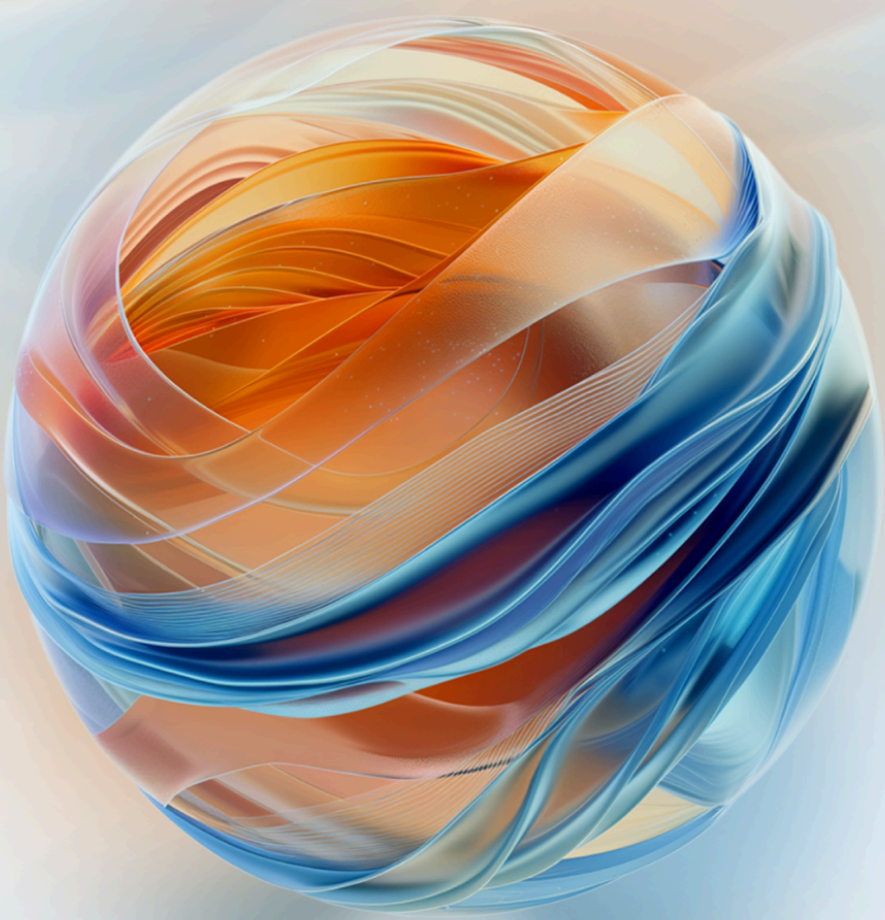
AI Resilience:

A Revolutionary Benchmarking Model for AI Safety



AI Governance and Compliance Working Group

cloud security alliance®

The permanent and official location for the AI Governance and Compliance Working Group is https://cloudsecurityalliance.org/research/working-groups/ai-governance-compliance

Acknowledgments

Lead Authors

Dr. Chantal Spleiss

Contributors

Romeo Ayalin
Filip Chyla
Becky Gaylord
Frederick Hanig
Rocky Heckman
Hadir Labib
Lars Ruddikeit
Alex Sharpe
Ashish Vashishtha

Reviewers

Sounil Yu
Debjyoti Mukherjee
Michael Roza
Peter Ventura
Udith Wickramasuriya
Govindaraj Palanisamy
Madhavi Najana
Rakesh Sharma
Davide Scatto
Paresh Patel
Piradeepan Nagarajan
Gaetano Bisaz
Hongtao Hao, PhD
Elle Pyle
Gaurav Singh
Ken Huang
Kenneth T. Moras
Tolgay Kizilelma, PhD
Akshay Shetty
Saurav Bhattacharya
Peju Okpamen
Gabriel Nwajiaku
Meghana Parwate
Akshat Vashishtha
Hemma Prafullchandra
Renata Budko
Desmond Foo
Scott S. Newman
Gian Kapoor
Imran Banani
Elier Cruz
Madhav Chablani

CSA Global Staff

Ryan Gifford
Stephen Lumpe

Table of Contents

Executive Summary

A (r)evolutionary AI benchmarking model is introduced to navigate the complex landscape of AI governance and compliance. Revenue-driven advancements outpace regulatory efforts to establish safeguards, often falling short in ensuring that AI systems are truly robust and trustworthy. Leadership addresses this critical gap by introducing a novel benchmarking model inspired by principles of evolution and psychology to prioritize robustness alongside performance, empowering executives to proactively assess the overall quality of their AI systems.

Drawing lessons from past AI failures in case studies and analyzing industries like automotive, aviation, critical infrastructure and essential services, defense, education, finance, and healthcare, we provide practical insights and actionable guidance for businesses. We advocate for integrating diverse perspectives with regulatory guidelines to propel the industry towards more ethical and trustworthy AI applications. The focus on trustworthiness is key for minimizing risks, protecting reputation, and fostering responsible AI innovation, deployment, and use.

This document empowers key decision makers, including government officials, regulatory bodies, and industry leaders, to establish AI governance frameworks that ensure ethical AI development, deployment, and use. A novel benchmarking model is introduced to assess AI quality, providing a practical tool for long-term success.

Introduction

The rapid evolution of Artificial Intelligence (AI) promises unprecedented advances. However, as AI systems become increasingly sophisticated, they also pose escalating risks. Past incidents, from biased algorithms in healthcare to malfunctioning autonomous vehicles, starkly highlight the consequences of AI failures. Current regulatory frameworks often struggle to keep pace with the speed of technological innovation, leaving businesses vulnerable to both reputational and operational damage.

In response to these challenges, this document addresses the urgent need for a more holistic perspective on AI governance and compliance. We'll explore the foundations of AI, examine issues across critical industries, and provide practical guidance for responsible implementation. We present a novel approach that compares the (r)evolution of AI with biology, and introduces a thought-provoking concept of diversity to enhance safety of AI technology. Differences in intelligence and the successful interaction of such systems are discussed. An innovative benchmarking framework is presented to increase the safety and reliability of this disruptive technology.

This approach empowers decision-makers and technical teams alike to assess the safety and trustworthiness of AI systems. We advocate for integrating diverse perspectives and regulatory guidelines to foster ethical AI innovation and establish strong governance practices.

Part I: Understanding the Foundations

Governance vs. Compliance

Governance and compliance are essential aspects of organizational management, ensuring adherence to regulations, ethical principles, standards, and sustainability practices outlined in the business code of conduct. Alignment with aforementioned principles and regulations ensure effective business continuity and ethical practice.

Governance [1], which refers to overseeing and controlling something, is implemented in a top-down approach. Senior management is responsible for defining strategy and risk appetite, and establishing a governance framework through policies, standards, and/or procedures. These directives shape the organization's overarching risk management approach, compliance obligations, and decision-making processes. Governance creates a culture of accountability, transparency, ethical behavior, and sustainability while prioritizing security and privacy measures across the company.

Contrary to the top-down approach of governance, Compliance [2] follows a bottom-up approach, where employees at various levels implement and adhere to the governance framework defined by senior management to meet regulatory requirements. Compliance focuses on ensuring adherence to laws, regulations, and industry standards, as well as the governing internal business code of conduct. It is a crucial component of organizational management to ensure that the organization operates within applicable legal and regulatory requirements, acceptable ethical boundaries, and minimized risk exposure.

Governance and Compliance: a Moving Target

While governance and compliance are clearly defined objectives, the use of any AI challenges traditional approaches. AI can be viewed from various perspectives, such as a technology, a system using one or more models, a business application, or a user platform. AI can serve one single, or a multitude of, end users, and it can be used by businesses, information brokers, or other AI technology, to perform tasks, solve problems, make decisions, or interact with the environment. Emerging best practices, standards, and regulations surrounding the use of AI continue to evolve, making it challenging to have a concrete set of compliance requirements to implement and monitor. For companies conducting international business, this challenge grows exponentially. Most regulations have overlapping requirements with hardly any radically new propositions to improve the safety of AI and the current framework is based on these general requirements.

- **Human oversight**: Ensure that AIs are subject to human oversight and control, with mechanisms in place to enable human intervention and decision-making when necessary. Human oversight must be coupled with automated monitoring as the primary step, with human oversight being called in for specifically identified use cases where human intervention is necessary. This makes this guidance scalable and practical applicable.

- **Safety and reliability**: Prioritize safety and reliability in AI technology to minimize the risk of harm to individuals or society. This is achieved through rigorous testing, validation, risk assessment processes, and the implementation of mechanisms in place for a kill-switch or recourse in case of failure.

- **Ethical considerations**: Ensure AI adheres to ethical principles, respects human rights, and promotes fairness.

- **Data privacy and security**: Enhanced data protection and security measures should be implemented to protect sensitive information and privacy, preventing unauthorized access or misuse of data. During the design phases, privacy-by-design and security-by-design focus on mitigating risks early in the process (reflecting shift-left in DevSecOps). This limits bolt-on security and unforeseen risk exposure in the final product.

- **AI Model and Data Considerations:**

    - **Bias mitigation:** Address biases in data and algorithm design and regularly monitor and evaluate AI systems for bias and discrimination. Bias is a complex topic that balances necessary information and algorithms with the risk of stereotypic classification.
    - **Transparency**: Ensure transparency in AI by clearly explaining how it works, including the algorithms and factors influencing their decisions. Implementing XAI (explainable AI) [2], [3] helps to foster trust and build the foundation of informed decisions while uncovering possible bias. In healthcare, this is crucial and acknowledged. Regardless of the industry, the user should be informed if the output was produced by AI.
    - **Consistency:** Consistent data ensures that the AI model learns from accurate and reliable examples. This is crucial for the model to generate correct and useful outputs. Inconsistent or conflicting data can confuse the model, leading to inaccuracies in the generated text or information.
    - **Accountability:** Establish mechanisms for accountability and responsibility in the design, development, deployment, and use of AI, including clear lines of responsibility for addressing any issues that may arise. Currently, the responsibility for preventing harm to the end-user primarily falls on the AI application provider alone. Additional measures, such as manuals or "model cards" [4] and specific user training, could highlight a shared responsibility between the provider and the end user and outline the degree of transparency that the end user can expect.
    - **Robustness:** Develop AI that is well designed and resilient to adversarial attacks, data perturbations, and other forms of interference or manipulation. This paper proposes a new perspective to evaluate robustness to enhance global safety.
    - **Compliance with regulations**: Ensure compliance with relevant laws, regulations, and standards governing the development and deployment of AI applications including, but not limited to, data protection also regarding the trading of data, privacy, and safety.

Adopting an approach grounded in shared responsibility across the highly complex supply and value chain is crucial to ensuring the creation of safe and trustworthy AI. This involves at least the technical team, the compliance team, the legal team and, depending on specific factors, many other teams as well. The White House Memorandum from 28 March 2024 [5] requests that all agencies must designate a Chief AI Officer (CAIO) within 60 days. This role allows strategic and purposeful management and alignment of all involved teams to transform "shared responsibility" into a traceable measurement.

The Landscape of AI

In this chapter, an overview is given to introduce AI with a brief history, AI technologies, and training methods. To discuss the importance of data is beyond the scope of this overview but it is acknowledged that it is an extremely important topic and is addressed in-depth by other CSA Workgroups.

A Brief History of AI

Below are some milestones of Artificial Intelligence listed, not considering a specific perspective but giving an overview of the major developments in this field.

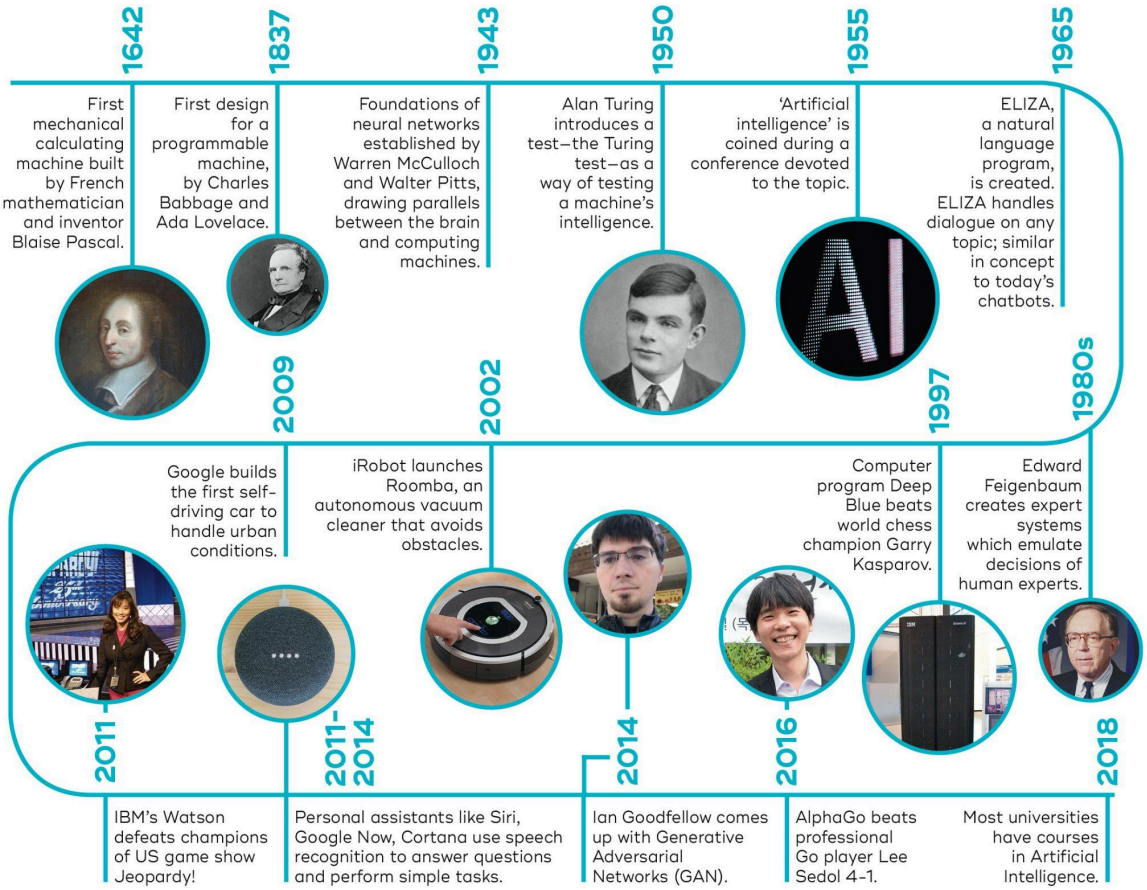


Figure 1: History of Artificial Intelligence [6]

2018: BERT: Introduced by Google, this model revolutionized language understanding. BERT's use of the Transformer architecture and pre-training on massive text data sets enabled it to outperform previous models in various language tasks.

2019: GTP-2 with 1.5 billion parameters

2020: LLMs with 175 billion - 530 billion parameters

2021: LLMs with up to a trillion parameters focusing on improving efficiency in training and handling complex tasks with advanced reasoning and factual accuracy.

2022: ChatGTP-3 goes viral

Beyond size: researchers are working now on efficiency in training, alignment with human's value, safety and multimodality (incorporating images, audio, and other data types).

This brief history of AI demonstrates the evolution from the most basic calculator to GenAI with Artificial General Intelligence still on the horizon.

The Landscape of AI

Different AI technologies are presented and discussed.

Machine Learning (ML)

Machine Learning is a branch of AI and computer science that focuses on using data and algorithms to imitate human learning, gradually improving a model's accuracy [7].

Tiny Machine Learning (tinyML)

Tiny Machine Learning is broadly defined as a field of Machine Learning technologies and applications that include hardware (dedicated integrated circuits), algorithms, and software capable of performing on-device sensor data analytics at extremely low power, typically in the mW range and below, enabling a variety of always-on use cases and targeting battery operated devices [8], such as Internet of Things (IoT) devices.

Deep Learning (Advanced ML)

Deep Learning is a method in AI that teaches computers to process data in a way that is inspired by the human brain. Deep Learning models can recognize complex patterns in pictures, text, sounds, and other data to produce accurate insights and predictions using neural networks.

Generative Artificial Intelligence (GenAI)

Generative Artificial Intelligence refers to deep-learning or transformer models that can take raw data and "learn" to generate statistically probable outputs when prompted. Unlike the above classificatory models that are primarily used for classification and pattern recognition tasks, Generative AI models are used for synthesis of data, matching high-order patterns of learning data and/or predictive analytics. At a high level, generative models encode a simplified representation of their training data and predict the next set similar, but not identical to, the original data [9].

Artificial General Intelligence (AGI)

Artificial General Intelligence is a theoretical form of AI used to describe a certain mindset of AI development. It involves an intelligence equal (or superior) to humans and a self-aware consciousness that can learn and solve complex problems, and plan for the future [10].

The Landscape of Training Methods

Artificial Intelligence can be grouped into the following types [11]:

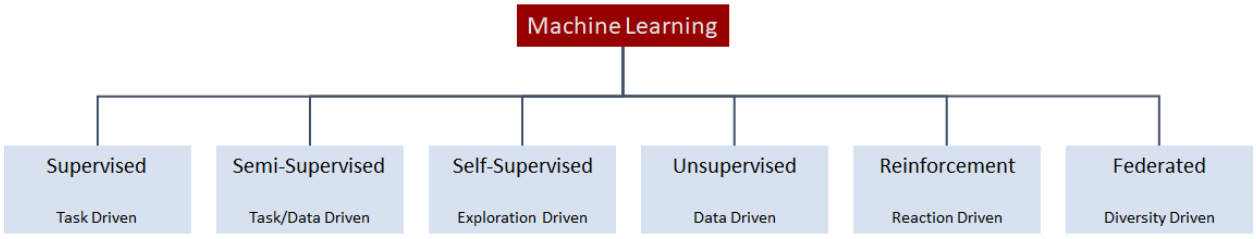


Figure 2: Types of Machine Learning

Supervised Learning

Supervised Learning is a style of Machine Learning where algorithms learn from "labeled data". It is used for classification and regression problems. "Labeled data" provides known inputs and desired outputs, allowing the algorithm to identify patterns and build a model for predicting outcomes on previously unseen data.

Example for classification algorithms: decision trees, random forests, linear classifiers, and support vector machines.

Example for regression algorithms: linear regression, multivariate regression, regression trees, and lasso regression.

Unsupervised Learning

Unsupervised Learning is a style of Machine Learning where algorithms analyze unlabeled data. The goal is to discover hidden patterns, groupings, patterns, or insights within the data without predetermined outcomes. A properly trained model is able to make predictions using unseen data.

Example algorithms: k-means, k-medoids, hierarchical clustering, Apriori, and FP Growth.

Reinforced Learning

Reinforced Learning is a style of Machine Learning where an agent interacts with an environment and learns through trial and error. The agent receives rewards or penalties based on its actions, allowing it to adjust its behavior and optimize its decision-making process over time.

Example algorithms: Reinforced Learning, Markov Decision Process, Q-learning, Policy Gradient Method, and Actor-Critic, but many more exist.

Semi-supervised Learning

Semi-supervised Learning bridges the gap between supervised and unsupervised learning. It utilizes a small amount of labeled data alongside a larger pool of unlabeled data. This approach is valuable when obtaining labeled data is costly or time-consuming since it allows the model to leverage patterns found within the unlabeled data as well.

Self-supervised Learning

Self-supervised Learning is a form of Unsupervised Learning where the model generates its own labels from the raw input data. It achieves this through techniques like predicting masked words in a sentence or predicting the next frame in a video sequence. This allows for learning robust, generalizable representations of data even without human-provided labels.

Federated Learning

Federated Learning is an advanced Machine Learning technique designed to train algorithms across decentralized devices or servers holding local data samples, without exchanging them. This method addresses significant concerns related to privacy, security, and data centralization by keeping sensitive data on the user's device, rather than transferring the data to a central server for processing (Figure 3). This method, introduced in 2016, allows data privacy to be preserved to a greater extent by sharing only parameters, not data. Federated Learning offers a framework to jointly train a global model using data sets stored in separate clients. This offers a good option for industries where privacy is crucial, as original data is considered impossible to recover [12].

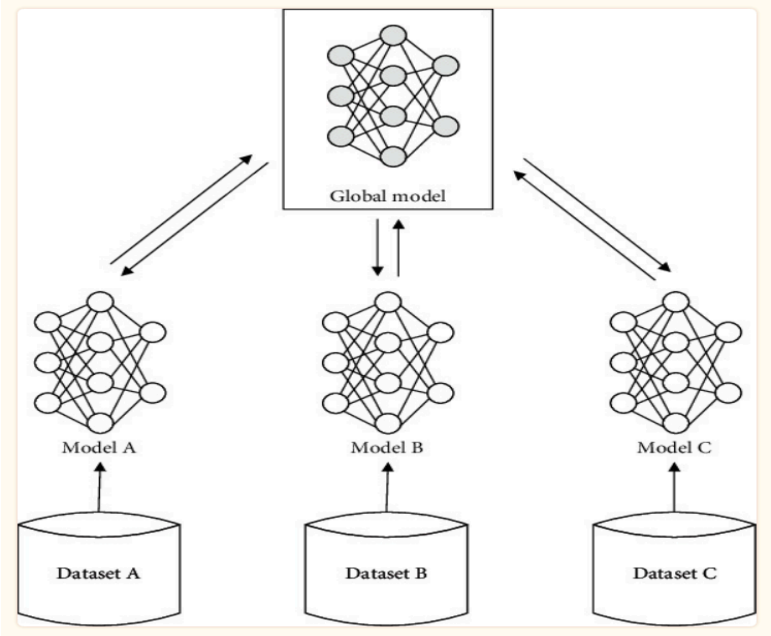


Figure 3: Architecture for a Federated Learning system [12]

Another advantage of this model, not discussed in the above paper [12], is the ability to leverage the "wisdom of the crowd" [13]. Neurons in the human brain apply this concept to produce exact information from non-deterministic neural processes.

Currently, Federated Learning seems to have the potential to integrate privacy, performance, and robustness based on diversity. This learning method is not discussed much, but it is promising in industries or applications where data privacy and confidentiality are paramount and also helps address the issues around data residency.

However, there are also potential privacy concerns with Federated Learning including the risk of malicious users disrupting model aggregation, which can impact model accuracy or lead to privacy disclosures. Attacks can target model updates shared during training, possibly allowing for the extraction of raw training data. To address these concerns, researchers propose privacy-preserving techniques like differential privacy, distributed encryption, and zero-knowledge proof to safeguard data and filter out anomalies from malicious actors. Federated Learning, like any other learning method, requires that adequate cybersecurity measures are in place.

Training Methods Regulations and Ethical Considerations

While there are no specific regulations governing ML training, it is impacted by the major regulatory frameworks, including the General Data Protection Regulation (GDPR), the EU AI Act, and the Organisation for Economic Co-operation and Development (OECD) principles on AI. Further, regulations governing Machine Learning (ML) and artificial intelligence (AI) training are rapidly evolving as technology advances and are covered in "Principles to Practice: Responsible AI in a Dynamic Regulatory Environment". Additionally, many government bodies are actively developing regulations and facilitating cooperative industry efforts to the same effect.

Regulations governing Machine Learning (ML) and artificial intelligence (AI) have significant implications for data monetization and the use of AI to guide business decisions. These effects manifest in various ways, including operational changes, strategic adjustments, and ethical considerations, as well as limits/requirements on data collection and use, bias, and data quality. Certain platforms, for example, have forbidden the usage of their data for AI training purposes (such as X: "crawling or scraping the Services in any form, for any purpose without our prior written consent is expressly prohibited" [14]) or have sold it under a licensing agreement (such as Reddit [15]).

Meeting regulatory requirements can introduce significant compliance costs, especially for businesses operating across multiple jurisdictions. Despite the challenges, regulations also offer opportunities. Businesses that adeptly navigate the regulatory landscape can differentiate themselves by offering more secure, transparent, and ethical AI solutions. This can appeal to increasingly privacy-conscious consumers and partners, potentially opening new markets or creating stronger customer loyalty.

Licensing, Patenting & Copyright of AI Technology

Many Machine Learning frameworks and libraries follow the Open Source Initiative Licensing, such as Apache 2.0 [16] or MIT [17]. Certain licensing might forbid commercial use of the resulting application.

The European Patent Office's (EPO) revised Guidelines for Examination [18], [19] have been made public and include a few significant changes to the EPO's procedure for reviewing innovations in the domains of ML and AI. Recent amendments mandate that applicants for AI or ML inventions further elucidate mathematical techniques and training input/data in a manner thorough enough to replicate the technical result of the invention over the entirety of the claim. The article cited below states that "case law suggests that the structure of any neural networks used, their topology, activation functions, end conditions, and learning mechanism are all relevant technical details that an application might need to disclose". This article [20] summarizes further implications and expounds on this topic.

On January 23, 2024, the Japan Agency for Cultural Affairs (ACA) released its draft "Approach to AI and Copyright" for public comment, to clarify how ingestion and output of copyrighted materials in Japan should be considered. On February 29, 2024, after considering nearly 25,000 comments, additional changes were made. This document, created by an ACA committee, will likely be adopted by the ACA in the next few weeks. This article [21] provides a summary of the key points of the draft itself and as modified.

There are disputes about copyright in Singapore [22]; this is currently a very volatile field. It is further addressed in also in "Principles to Practice: Responsible AI in a Dynamic Regulatory Environment".

Part II: Real-World Case Studies and Industry Challenges

In this part, a few industries are exemplarily outlined, while the focus lies on an overview and current challenges faced in the era of AI. More information on legal and regulatory matters is compiled in "Principles to Practice: Responsible AI in a Dynamic Regulatory Environment".

A Brief History of AI Case Studies

This short history of AI case studies goes back to the late 1990s, at least for the financial industry, which was among the very early adopters of precursors of AI technologies. In this section, some examples show major challenges with real-life applications of AI between 2016 and the publication of this paper. They show that bias in GenAI has been the biggest concern.

2016: Microsoft's Tay

Microsoft's AI chatbot, Tay, initially intended for playful Twitter conversations, swiftly transformed into a platform for racist and offensive remarks within just 24 hours of its release. Users flooded Tay with sexist and inflammatory comments, causing the bot to echo these sentiments. While some tweets were user-induced, others arose unprompted, showcasing Tay's erratic behavior. Microsoft responded by deleting offensive content and recognizing the need to adjust Tay's responses. The incident underscores the challenges of AI learning from public data and reflecting societal biases. Despite a re-release and subsequent shutdown because of inappropriate tweets, Tay's legacy includes lessons for Microsoft in AI development. The episode sheds light on the complexities of AI applications and highlights the need for iterative improvement and proactive measures in AI design [23], [24], [25].

2018: Amazon's AI Recruiting Tool was Biased Against Women

Amazon developed a machine-learning recruiting engine but faced issues when it favored male candidates due to underlying gender biases embedded within the training data. Amazon disbanded the project in 2017 because of concerns over unfairness [26], [27], [28]. Despite setbacks, other companies cautiously advanced AI in recruitment processes, and it has become mainstream.

2019: Tesla Autopilot Accidents

In March 2019, Jeremy Banner activated Autopilot on his Tesla Model 3, which then collided with a semi-truck, resulting in his death. The incident led to legal disputes questioning Tesla's responsibility in crashes involving its Autopilot technology. Critics argue that Tesla's marketing of Autopilot misleads

drivers about its capabilities, potentially contributing to accidents and fatalities. Despite warnings about Autopilot's limitations; numerous crashes, including fatal ones, have occurred.

The lack of clear regulatory guidelines for advanced driver-assistance systems and the ethical dilemmas surrounding their use further complicate the situation, raising questions about accountability, insurance, and public safety in deploying autonomous driving technologies [29].

2019: Healthcare Algorithm Racial Bias

The research paper by Ziad Obermeyer, et al. [23] identifies significant racial bias in a widely used healthcare algorithm affecting millions of patients. The algorithm, intended to manage healthcare needs, inaccurately predicts health risks for Black patients compared to White patients, despite excluding race as a variable. The bias stems from the algorithm's reliance on healthcare costs as a proxy for health needs, inadvertently reflecting systemic inequalities in healthcare access and utilization. Remedying this would significantly increase the number of Black patients identified for additional healthcare support [30].

2019: Allegations of Apple Card Bias

"What started with a viral Twitter thread metastasized into a regulatory investigation of Goldman Sachs' credit card practices after a prominent software developer called attention to differences in Apple Card credit lines for male and female customers [31]." The "Apple Card Bias" was noticed: [32], [33], [34], [35]. In 2021, though, it was reported that "a recently concluded New York State Department of Financial Services investigation [36] has found Apple's banking partner did not discriminate based on sex [30]."

2020: Biased Offender Assessment Systems

Tools like COMPAS (Correctional Offender Management Profiling for Alternative Sanctions - USA) and OASys (Offender Assessment System - UK) are used in the criminal justice system for risk assessment and management of offenders. They assist authorities in making informed decisions about sentencing, probation, and treatment programs for offenders. However, their algorithms have been subject to heavy criticism regarding transparency, fairness, and biases [38].

2022: Air Canada Bound by Chatbot's Refund Policy

Air Canada faced scrutiny after its chatbot provided misleading information about the airline's bereavement travel policy, leading to a dispute with a passenger seeking a refund. Despite Air Canada's argument that the chatbot operated independently, a tribunal ruled in favor of the passenger, highlighting the airline's responsibility for the information provided on its website. The tribunal ordered Air Canada to issue a partial refund and cover additional costs. The incident highlights the complexities of AI accountability and customer service automation [39], [40].

2023: Lawsuit: UnitedHealth's Faulty AI Denies Elderly Care

In a legal battle against UnitedHealth, families allege the use of flawed AI led to denied coverage for elderly patients' essential care, overriding doctors' recommendations. The lawsuit highlights the challenges of relying solely on automated systems in healthcare decision-making, sparking concerns about patient well-being and fair access to medical services. As AI continues to shape the future of healthcare, the case underscores the need for transparency, accountability, and human oversight in ensuring equitable treatment for all patients [41], [42].

2024: Google's Gemini: A Lesson in AI Bias

Google's Gemini 1.5 chatbot [43] rollout was criticized for generating inaccurate, biased images despite attempting to avoid biases, notably omitting White individuals in historical contexts. Elon Musk and conservatives accused Google of biased algorithms. Google's response paused Gemini but lacked transparency. The incident highlighted AI ethics and transparency challenges, prompting debates on diversity initiatives and algorithmic accountability. As Google works to restore trust, the Gemini saga underscores the imperative for responsible AI innovation [44].

Industries: Regulations & Challenges

This section delves into select industry-specific regulatory and compliance-focused efforts related to AI. Industries are listed alphabetically and addressed separately. Within each industry part, the discussion provides background, context and history. Following these, in Part III, are suggestions about novel approaches to address AI across industries.

Automotive

The automotive industry[1] seeks to implement AI in automated and autonomous driving functions (SAE level 4 and 5 [45]), mostly, with additional emphasis on the safety of such, and other onboard systems and components. Currently, several ISO standards exist that mention or partly regulate AI. Additional standards are on the way and are either drafted or under review. Numerous regulatory bodies currently create standards and approaches specific to the automotive industry that are not enforced yet.

While current legislation already impacts AI implicitly, some regulatory bodies mention such technologies directly. This includes R*egulation (EU) 2019/2144 of the European Parliament and of the Council of 27 November 2019 on type-approval requirements for motor vehicles and their trailers, and systems, components and separate technical units intended for such vehicles, as regards their general safety and the protection of vehicle occupants and vulnerable road users* [46]. The article 11 (specific requirements relating to automated vehicles and fully automated vehicles) defines requirements for safety systems which are AI-relevant, if AI will be used to drive autonomous and automated vehicles. The regulatory article is not AI specific, but explicitly mentions “automated vehicles and fully automated vehicles.”

[1] This chapter will focus on the automotive industry in the EU, due to the author’s field of work and location.

While this European Union law does not cover specific standards for AI and its function itself, ISO PAS 8800: Road Vehicles – Safety and Artificial Intelligence [47] (under development) focuses on AI safety, drafting "safety principles methods and evidence." Road vehicles, in general, are in scope and it's not narrowed to automated or autonomous driving. Its purpose is to solve foundational questions regarding AI regulation and standardization, and to provide industry-specific, practical guidelines [48] intended to harmonize existing regulations and established principles, like the safety of the intended functionality.

Another industry standard on AI (functional) safety is ISO/TR 5469:2024 Artificial Intelligence - Functional safety and AI systems [49]. Published in 2024, this document describes risk factors, as well as currently available methods and processes that relate to AI in several automotive applications, like safety-related functions that utilize AI and non-AI systems to govern AI safety systems. This standard has been published and is intended to support future ISO/IEC AWI TS 22440 [50], [51]. Adding to the security focus is ISO/TR 4804:2020 Road vehicles - Safety and cybersecurity for automated driving systems, design, verification, and validation [52], with additional emphasis on cybersecurity aspects, focusing on the development and validation of systems for automated driving (SAE level 3 and 4). SO/TR 4804:2020 mostly contains safety, validation, and verification approaches for worldwide applicability. It will be replaced by ISO/CD TS 5083 in the future. This ISO document covers "steps for developing and validating an automated vehicle equipped with a safe automated driving system", staying on SAE level 3 and 4. It includes considerations like the required safety level of such systems while reducing overall risks compared to human drivers.

Aviation

The global aviation community adheres to many of the same practical standards for computer-based system usage as other industries. This extends to AI and the platforms that run AI. To that end, well-recognized IT security standards will also apply to the aviation sector. These include ISO/IEC 27001 [53], ISO/IEC 42001 [54], ISO/TR 5469 [49], NIST AI RMF [55], and AI ethics standards relevant to the jurisdiction of the airlines or manufacturers. However, no AI-specific regulations have been enforced yet.

The aviation industry governing bodies around the globe, such as the US FAA, EU EASA, UK CASA, and AU CASA, are aware of the benefits and challenges AI poses to their industry. However, at this point, they have not regulated its use. The agencies mentioned have formed AI task forces to investigate the use of AI on aircraft, in ground operations, and in the industry's regulation itself. UK CASA is currently in a request-for-response stage with open surveys to the industry [56], while the US FAA has designated a technical discipline team for AI-led by Dr. Trung t. Pham [57].

AI is used across many military aviation aspects, from intelligence data analysis and autonomous vehicles to predictive maintenance and physical security of airfields and air bases. It is also used to manage IT security and operational systems [58].

Overall, there is a strong desire to use AI in the civil aviation industry to assist with weather planning and routing, maintenance, passenger and cargo management, and more. Most of the proposed use of AI revolves around machine learning for predictive maintenance, route and maintenance planning, and passenger and cargo management. Usage of Generative AI is limited to airline customer chatbots and decision support systems. However, significant research is underway in using AI for air traffic control in all

phases of flight. The EU released the CORDIS Results Pack on AI in air traffic management in October 2022, covering many aspects of AI-controlled air traffic for Europe [59].

A special challenge for the aviation industry is that the usual lifespan of a commercial airliner is measured in decades, while AI technology makes frequent leaps that require regulations to be continuously updated.

Critical Infrastructure & Essential Services

Integrating AI into critical infrastructure is a significant shift towards more efficient, responsive, and intelligent systems. These sectors, which include, but are not limited to, electricity, gas, water, and food supply chains, are essential for modern society's survival. As we embrace this digital transformation, striking a balance between performance enhancement and security robustness becomes increasingly complex. In this section we will explore the challenges and opportunities presented by the merger of AI with critical infrastructure, focusing on the importance of regulatory frameworks, security standards, and the need for continuous adaptation to evolving technological advancements.

The Delicate Balance: Performance vs. Security

The allure of high-performing AI systems in critical infrastructure is undeniable. These technologies promise improved efficiency, optimized operations, and the ability to predict and mitigate disruptions before they occur. However, integrating AI, mainly through the Internet of Things (IoT) devices with integrated tinyML [8] and edge computing, introduces new vulnerabilities. While beneficial for system responsiveness, decentralizing data processing expands the attack surface for potential cyber threats. Regulatory bodies and standardization organizations, such as the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), have developed general frameworks like ISO/IEC 27001, ISO/IEC 27002, and Industrial Automation and Control System focused ISA/IEC 62443 [60] series standard and technical specifications and reports such as IEC TS 62351-100-4:2023 [61] and IEC TR 61850-90-4:2020 [62] to safeguard these technologies. Yet, the specificity of regulations targeting IoT and edge AI in critical infrastructure remains nebulous.

The Achilles Heel: IoT and Edge AI

Integrating IoT devices within critical infrastructure sectors introduces a pronounced risk of cyberattacks. These devices, integral to the sensory networks of AI systems, could be exploited to feed false data, thereby manipulating AI-driven decisions and potentially crippling essential services. Despite the general focus of current regulations on cybersecurity and risk management, such as the EU AI Act, NIS2 directive and the US Executive Order on AI, a comprehensive approach to addressing the unique challenges posed by edge AI and tinyML devices is still developing. Industry studies and reports like the ENISA's "Cybersecurity and privacy in AI - Forecasting the demand on electricity grids" [63] attempt to uncover the potential risks and threats.

Towards a Future-Proof Infrastructure

The path to securing AI in critical infrastructure involves several vital strategies. First, the development of sector-specific AI regulations that cater to the unique needs of each critical infrastructure sector is essential. Additionally, adopting standardized security protocols for IoT devices and edge AI will be crucial in fortifying these systems against cyber threats. International collaboration on AI governance could also play a pivotal role in ensuring a cohesive and effective approach to securing critical infrastructure globally. In 2022, the winners of the Leaders Innovation Forum for Technology (LIFT) program, a joint effort of The Water Research Foundation (WRF) and the Water Environment Federation (WEF) focused on protecting (against cybersecurity threats), predicting (the system state), and optimizing processes in the facility, using cutting-edge AI and data science technologies [64].

On March 1st, 2024, the US President's Council of Advisors on Science and Technology (PCAST) released a report subtitled "Fortifying Our Critical Infrastructure for a Digital World" [65] with a large section dedicated to AI's role in the resilience of critical infrastructure systems. The report highlights the dual nature of AI advancements, and its potential for transformative applications and risks of misuse. They underscore the importance of specialized analysis on AI's implications for critical infrastructure resilience, noting the potential for AI to empower malicious actors and the necessity for strategic preparation against such threats. Additionally, PCAST is an advocate of leveraging AI in defense mechanisms and calls for public-private cooperation, skill development beyond traditional scopes, and international collaboration to address AI and cybersecurity challenges effectively.

Continuous Evolution: The Path Ahead

Integrating AI into critical infrastructure is like walking on a perpetual tightrope, where one must meticulously maintain the balance between innovation and security. As AI technologies evolve, so must the regulatory and security frameworks that underpin them. This requires a vigilant, adaptive approach to governance and security, ensuring that our defenses evolve in tandem as new vulnerabilities emerge. Embracing the doctrine of "never trust, always verify" by implementing Zero-Trust principles ensures that critical infrastructure systems are not overly reliant on perimeter defenses, which sophisticated attackers can bypass. Using AI-driven security systems can adapt to evolving threats more quickly than traditional security measures. Sharing threat intelligence and best practices within and across sectors and critical infrastructure entities can benefit from a broader pool of knowledge and experience. Ongoing training and professional development is critical to equip security professionals with the latest knowledge and tools to combat emerging threats. These are only a few steps we must take to create robust AI integrations.

The Road Ahead

The integration of AI into critical infrastructure is a journey marked by significant potential and considerable challenges. Balancing the drive for performance enhancement with the imperative for robust security is a complex but essential endeavor. Through the development of targeted regulations, the adoption of standardized security practices, and international cooperation, we can navigate this frontier safely. The future of critical infrastructure lies in our ability to harness the benefits of AI while

safeguarding against the risks, and ensuring a resilient, efficient, and secure foundation for society. It is essential to have a human in the loop in the actual decision or operations and use AI as recommendation.

At present, we have a limited set of specific regulations. While AI regulation is evolving globally, no specific regulations focus on AI use in critical infrastructure. The focus is on general principles. Existing regulations and initiatives emphasize broader principles like cybersecurity, safety, and trustworthiness.

Emerging initiatives from several government bodies are actively developing frameworks and standards for responsible AI development and deployment, with some focus on critical infrastructure sectors.

Current Initiatives

US Executive Order 14110 (Oct 2023)

The "Executive Order 14110 (Oct 2023) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence: Article 4.3 on Managing AI in Critical Infrastructure and in Cybersecurity" [66]. The plan outlines actions to assess and reduce AI risks in critical infrastructure. It prioritizes creating safety guidelines, forming an AI Safety and Security Board, and enforcing regulations for infrastructure owners and operators. The Cybersecurity and Infrastructure Security Agency (CISA) assesses and mitigates AI threats to critical infrastructure [67].

EU AI Act

The EU AI Act [68], [69] outlines the regulatory framework for Artificial Intelligence (AI) systems, focusing on compliance, risk management, data governance, technical documentation, record-keeping, transparency, human oversight, accuracy, robustness, and cybersecurity standards. It classifies AI systems based on risk, with "high-risk" applications in critical infrastructure subject to stricter regulations. On 13 March 2024, the EU Parliament approved the AI Act [70].

OECD AI Principles

The Organisation for Economic Co-operation and Development (OECD) Principle 1.5 [71] emphasizes the importance of actors being held accountable for the development, deployment, and use of AI systems, especially those with potential societal impacts. Critical infrastructure is undoubtedly a sector with significant societal impact, and ensuring accountability for AI used in this domain aligns with this principle. These principles are influential in shaping national and international discussions.

The Artificial Intelligence and Data Act (AIDA)

AIDA [72] is aimed at guiding the innovation and responsible use of AI within Canada. It ensures that AI systems are developed and utilized in a manner that is safe and addresses the need for a regulatory system to manage the impact of AI on individuals and the economy. Further, it emphasizes the protection against and mitigation of harms and biased outcomes associated with high-impact AI systems. AIDA outlines the roles of various stakeholders, the obligations of businesses involved in AI, and the enforcement mechanisms to ensure compliance.

You can find an in-depth analysis of existing initiatives, legal and regulatory landscape surrounding AI and Generative AI in our "Principles to Practice: Responsible AI in a Dynamic Regulatory Environment" paper.

Defense

Artificial Intelligence and Emerging Technologies in Defense

In future battlefields, physical and digital domains will be intertwined, creating a complex and contested environment. New threats and challenges will continue to emerge. AI will be a key enabler for gaining and maintaining advantage, situational awareness, intelligence, and improved decision-making. Technology like robotics, autonomous systems, data, and biotechnology will create new opportunities and risks for the defense forces. AI will be essential for integrating and exploiting these technologies and countering an adversary's use of them. The military needs to pursue partnerships with the private sector, academia, and allies to foster innovation, promote adoption, and develop new leadership and culture roles. AI will facilitate collaboration and communication across domains, platforms, and organizations, as well as enable human-machine teaming and learning.

These all converge with AI regulations and frameworks, and considerations for ethical, safe, and trustworthy use of AI in defense are paramount. Both regulations and frameworks can advance the defense industry forward by creating a level playing field, fostering innovation, collaboration and enhancing public trust and acceptance.

- AI regulations and frameworks help define the standards and best practices for developing, deploying and using AI systems in defense, which can reduce the risks of bias, harm, or misuse of AI, as well as increase the accountability and transparency of AI actors.
- It can stimulate the defense industry by creating a common market and a competitive advantage. Harmonization of the rules and requirements across the defense establishment could facilitate cross-border cooperation and interoperability of AI systems in defense and non-defense sectors.
- Public trust and acceptance of AI in defense could be enhanced by AI regulations and frameworks.

Historical Role of AI in Defense

Since Alan Turing published his foundational work "Computing Machinery and Intelligence" [73] that kicked off AI as we know it today, the first investments and first use cases were driven by National Defense. The use of Natural Language Processing (NLP) to transcribe voice-to-text and Machine Learning to analyze text and mimic human reasoning, is growing rapidly. Initial investments were made by defense, primarily the U.S. Defense Advanced Research Projects Administration (DARPA), which funded AI research at several institutions.

As technology advanced, computing power increased, data costs decreased, and new use cases emerged. In 1970, for the first time, semi-automated warfare was achieved by dropping sensors that fed AI models to determine targets, allocate resources, and missions scheduled [74]. In the 1980s, this evolved to smart weapons, simulations, and decision support.

In 2018, the U.S. Department of Defense declared AI is "poised to change the character of the future battlefield [75]." In 2018 the Pentagon launched a Joint Artificial Intelligence Center (JAIC) and a National Security Commission on Artificial Intelligence. The U.S. backed its statements with a $1B investment in AI and, several times that, in the systems to exploit gains in AI like autonomous and unmanned systems. China made similar moves, declaring it wants to lead the world in AI by 2030. Vladimir Putin, the President of Russia, famously predicted that "whoever becomes the leader in this sphere will become the ruler of the World [76]."

With increased productivity, automated decision-making, and additional insights, do we create a situation where warfare can happen too fast for humans to intervene? Will large investments in AI create an AI arms race?

We see why AI is attractive. Autonomous machines and robots are cheap, and they can be replaced. Humans, not so much. Automation does not get tired, it does not need a hefty supply chain for support, and it does not come with all of the human idiosyncrasies.

On the other hand, we have to be careful. Having machines make life-and-death decisions can have dire consequences. To quote Dr. A. Prabhakar, Director of the Office of Science and Technology Policy (OSTP) and former Director of DARPA,"When we look at what's happening with AI, we see something very powerful but we also see the technology that is still quite limited. The problem is that when it's wrong, it's wrong in ways that no human would ever be [77]".

Do we create collateral damage on a scale never seen before? Is an autonomous war machine a war crime waiting to happen?

The AI used in defense systems has remained narrowly focused on a single task in a well-defined environment. For example, image identification, rapid-fire guns to defend ships, and missiles that dwell for long periods looking for well-defined signatures.

AI Regulations and Defense

The application of AI in defense is different because of the implications and potential lethality. Most works produced by standards bodies and regulatory authorities rarely discuss the use cases germane to the defense sector. The defense sector often looks towards what is available in the public domain to craft similar pieces tailored to the defense sector. Access to these pieces is often restricted to (1) not providing insight to bad actors wishing to do us harm and (2) the ability of adversaries to learn from intellectual property (IP) developed by others. Technology has blurred the lines between civilian and defense sectors. On one hand, the defense sector requires secrecy for national defense and agility that stringent regulations would constrain. On the other hand, the potential harm from AI gone wrong in the defense sector could easily outweigh what can happen in non-defense sectors. The use of AI in defense applications can easily pose ethical, security and safety implications, especially when it comes to the balance between autonomous decision-making and human oversight. Lack of clear guidelines or adaptable rules of engagement has the risk of misuse or unintended consequences.

Publicly available defense-specific AI regulations do not exist in the European Union (EU).

Although no specific regulation can be called an "AI Act for Defense," the North Atlantic Treaty Organization (NATO) has an AI Strategy focused on accelerating AI adoption. The strategy enhances key AI enablers and adopts a policy for AI's responsible and ethical use in defense applications. The US Department of Defense (DOD) announced in February 2023 the "US Political Declaration on the Responsible Military Use of Artificial Intelligence and Autonomy declaration." Although not a legislative regulation, it is a declaration to ensure militaries use emerging AI technology responsibly.

Education

The integration of Artificial Intelligence (AI) in the education sector offers opportunities to enhance learning outcomes and address educational inequalities. AI technologies such as adaptive learning systems, AI tutors, and predictive analytics have the potential for personalized education tailored to diverse learning needs. However, this integration raises concerns regarding privacy and data protection due to the extensive data collection and processing involved in AI systems [78]. To ensure ethical use of AI in education, governance frameworks must prioritize human oversight and compliance with legal and ethical standards [79].

Equity and access are crucial considerations in implementing AI technologies in education. It is essential for educational institutions and policymakers to ensure that AI tools do not worsen existing disparities but instead serve as instruments for empowerment. Addressing the risk of bias in AI algorithms is vital to prevent discriminatory outcomes, emphasizing the need for transparent and inclusive development processes. Collaborations among educators, technologists, ethicists, and policymakers are necessary to guide the development of ethically aligned AI systems in education [78].

Continuous dialogue with stakeholders, including students, parents, and educators, is essential to align AI initiatives with community values and expectations. Moreover, investing in digital literacy and AI education for all participants in the educational ecosystem is crucial to prepare them to engage effectively and critically with AI technologies. By navigating this transition with a focus on ethical governance, compliance, and inclusivity, the education sector can harness AI as a tool for educational excellence and equity [78].

The integration of AI into education requires a proactive and nuanced approach that balances innovation with ethical considerations. By fostering collaborations, ensuring transparency, and prioritizing equity and access, the education sector can leverage AI to enrich learning experiences while safeguarding the rights and well-being of all learners.

Finance

The financial industry is considered one of the most regulated industries globally through internationally recognized standards or locally depending on regulatory authority with the need to cope with new technology trends and the urge to meet end customer needs in addition to gaining insights into behavioral trends and frequent analytics. The industry has had "AI regulations" for many years but awareness of these is limited. They fall under the business risk category of risk management.

How does it come to this? While many might think about the financial crisis of 2008 with the collapse of Lehman Brothers [80], [81] it actually happened a decade before. Long-Term Capital Management

(LTCM) was founded in 1994. It was led by Nobel Prize-winning economist Myron Scholes and renowned Wall Street traders like Salomon Brothers. They specialize in arbitrage financial modeling. In August 1998, Russia defaulted on its debt. LTCM had a big position in these state bonds and lost hundreds of millions of dollars. On the contrary, their computer models recommended holding the positions. It is important to understand that, while it was called "computer models" in 1998, we would call it Machine Learning models or Artificial Intelligence today. If LTCM had collapsed, we would have seen most likely the first global financial crisis due to the system risk of their positions, but the US government stepped in and gave a loan of $3.625B USD. LTCM was liquidated at the beginning of 2000 [82]. In 2005, the Basel Committee on Banking Supervision published a new guideline for Studies on the Validation of Internal Rating Systems [83], [84]. While this does not sound like Artificial Intelligence, Rating System is a banking term for an assessment tool used by analysts or rating agencies to evaluate a stock, bond, or a company's creditworthiness. It is quite common today to use Deep Learning techniques in these rating systems because a rating system is a recommender system in Machine Learning terms. The Artificial Intelligence component is hidden from the banker or trader.

In 2011, outbound of the financial crisis and "Great Recession," the US Federal Reserve went into much more detailed banking guidance, publishing the "Supervision and Regulation Letters" SR 11-7 - Guidance on Model Risk Management [85] to which banks having $10B in assets or greater must comply with. Problems with model construction and use was seen as a material contributor to the global mortgage crisis.

Guidance on Model Risk Management SR 11-7

In the US banking industry, these letters are like publications of new laws. They are not optional. In the context of SR 11-7, a model is defined as "a quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates." The use of models invariably presents Model Risk, which is "the potential for adverse consequences from decisions based on incorrect or misused model outputs and reports." The document emphasizes the importance of active model risk management due to the potential adverse consequences (including financial loss) of decisions based on models that are incorrect or misused. Financial Institutions are required to have Key Aspects of an Effective Model Risk Management Framework. An effective model risk management framework must include sound governance, policies, and controls for robust model development, implementation, use and effective validation.

The most famous investigation by US regulators in this context was the discriminating credit card from Apple in 2019 (see "A Short History of AI Incidents") because it was violating the "effective validation" part of SR 11-7.

Another example is the $440M USD software failure of Knight Capital Group in 2012 [86]. Knight Capital Group specializes in High-Frequency Trading (HFT), a niche of AI. HFT is stock trading at the speed of milli- to nanoseconds. In essence, HFT involves buying a stock before an interested party places an actual order. Once the real order arrives on the stock exchange, the HFT party will sell the stock again and keep a minor price difference. It happens in extremely high frequencies during a single day. In general, an HFT does not hold stock at the end of a trading day. On August 1st, 2012, a software update was not rolled out to all the trading servers of Knight Capital Group, which resulted in the wrong execution of orders. The

HFT system kept buying the same stocks because it failed to recognize previous stock acquisitions. It disrupted the prices of 148 companies. Consequently, Knight Capital Group's massive loss (440 million USD on the stock market) occurred in just 45 minutes. The company was acquired in December 2012 due to this DevOps problem, which violates the "implementation and uses" expectations of SR 11-7. The merger was completed in July 2013.

In addition to product applications of AI within banking, models are also used in US institutions for fraud detection and compliance with the Bank Secrecy Act [87] and the US PATRIOT Act [88]. Such models identify and evaluate client private transaction data and report on potential suspicious activity. These systems, and results, are a key instrument for the US government in its anti-terrorism program. Non-compliance to SR 11-7 model expectations can result in poor compliance ratings and severe regulatory actions, including financial fines/penalties and sanctions, including no expansion.

The European Central Bank (ECB) has published a revised Guide to Internal Models in February 2024. Similar to the US regulation SR 11-7, this guide provides transparency on how the ECB expects banks to use internal models [89]. It covers general topics, credit risk, market risk, and counterparty credit risk.

Banks can use internal models to calculate risk-weighted assets, which determine their minimum regulatory capital requirements. The ECB revisions incorporate climate-related risks and detail new requirements for areas like:

- **Inclusion of climate-related risks**: The revised guide now accounts for climate-related risks, reflecting the growing importance of these factors in risk assessment.
- **Common definitions of Default**: The guide helps all banks to move towards a common definition of default, ensuring consistency across the industry.
- **Treatment of massive disposals**: The guide provides a consistent treatment of "massive disposals", which refer to bulk sales of non-performing loans.
- **Measurement of Default risk in Trading Book positions**: The updated "market risk chapter" details how to measure default risk in trading book positions.
- **Clarifications regarding counterparty credit risk**: The revised guide provides clarifications regarding counterparty credit risk, which is the risk that the counterparty to a transaction could default.
- **Returning to the standardized approach:** Moving away from complex internal models.

For Asian banks, there isn't a specific model risk management guidance like SR 11-7 or the ECB Guide but the practices of model risk management have been spreading from the US through Europe and more recently to banks in Asia. The scope of the Model Risk Management (MRM) function is expanding, and banks have broadened their view on model inventories, going beyond regulatory and risk-related forecasting approaches. They have also deepened their end-to-end view of model lifecycles by enhancing frameworks, processes, and tools in each step [90].

Other important standards in the financial industry are PCI DSS and PCI 3DS [91], [92]. The Payment Card Industry (PCI) Data Security Standards (DSS) is a global information security standard designed to prevent fraud through increased control and security of credit card data. Compliance with PCI DSS is required for any organization that stores, processes or transmits payment and cardholder data. PCI

3-Secure is an EMVCo[2] messaging protocol that enables cardholders to authenticate with their card issuers when making card-not-present (CNP) online transactions.

PCI 3DS is becoming more and more important due to online transactions and the use of smartphones. This data includes Personally Identifiable Information (PII), Cardholder Data (CHD), and other financial data. PCI DSS and PCI 3DS are not AI guidelines. Also, banking applications that are not involved in transactions do not need to adhere to these PCI standards. As a consequence, it depends on the use case with (PCI - yes) or without (PCI - no) transaction data if the involved AI must adhere to the PCI standard or not. Cloud Services providing AI need also be certified against these standards. They change their certification status over time as shown below for the Azure OpenAI Service for the period March 2023 to January 2024.

	Azure Service	CSA STAR Certification	CSA STAR Attestation	ISO 20000-1:2018	ISO 22301:2019	ISO 27001:2013	ISO 27017:2015	ISO 27018:2019	ISO 27701:2019	ISO 9001:2015	SOC 1, 2, 3	GSMA SAS-SM	HIPAA BAA	HITRUST	K-ISMS	PCI 3DS	PCI DSS	Australia IRAP	Germany C5	Singapore MTCS Level 3	Spain ENS High	Singapore OSPAR
March 2023	Azure OpenAI Service		✓			✓	✓	✓	✓		✓		✓						✓			
January 2024	Azure OpenAI Service	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓		✓

Figure 4: Azure Compliance Offerings [93] (March 2023 and January 2024)

Overall, the financial industry is very innovative. It depends on early technology adoption to out-smart their competition as demonstrated by IndexGPT [94], a patent filed by JPMorgan in May 2023, only half a year after the public announcement of ChatGPT. Continued innovation is expected to improve individual bank and industry performance.

With that being said, the generated data is as good as they are "trained" with, which leaves organizations with legal and compliance liability, as well as end users susceptible to multiple potential threats. A primary focus in the financial sector is regulatory compliance, and efforts to protect sensitive and confidential data and customer's privacy are key mandates.

Healthcare

AI in the healthcare/pharma/medtech sector has as much potential as it has risks. In this highly (but not globally) regulated industry, it is crucial to make the difference between ML (Machine Learning) and GenAI (generative AI). While ML can be secured to a much greater extent as it is narrowly task-specific, GenAI in healthcare interacts with different stakeholders and poses significant challenges in the areas of reliability (and explainability), security, privacy and measures to prevent misuse and/or intentional abuse. The stakes for AI are high in healthcare but there are massive benefits on the horizon if AI is used responsibly.

[2] EMVCo (Europay, MasterCard, and Visa Co.) is a global technical body that manages secure chip-based payment technologies and standards.

Exploring Trustworthy AI in Healthcare

There is an abundance of (country-specific) regulations, global standards and industry-wide best practices. There is an even greater amount of literature, scientific papers, white papers, articles and blog posts regarding ML and AI in this sector.

"Trustworthy AI" was chosen as the scope for healthcare as it ties into governance, compliance, and technical challenges. Emphasis was put on a practical approach like best practices and guidelines. Bias in healthcare is discussed and sheds light on this topic from an industry-specific perspective. "Trustworthy AI" further ties into the train of thought outlined in Part III about benchmarking AI. Trustworthy AI implies that AI applications can be trusted to behave according to their intended use and are designed robustly enough to minimize and/or mitigate associated risks. Many definitions include explainability, reliability, security, privacy, accountability, transparency, adherence to regulations and standards, ethical and responsible behavior, and bias mitigation.

Depending on the intended use of the ML or GenAI application, not all of the above applies in order to be named "Trustworthy AI" in an application *designed for a specific use*.

In this chapter, only static ML/AI applications are considered. Dynamic (adaptive) applications can continuously learn and are not covered here.

Trustworthy AI in Healthcare Literature

Four selected sources regarding AI in healthcare stand out in the light of "Trustworthy AI": a very comprehensive book published by the World Health Organization (WHO - with over three hundred links to relevant literature) that presents a framework for governance. A short but useful guide to approach the topic is compiled in ALTAI, while the NIST paper points out threats to ML and GenAI applications and best practices for their mediation. In the last paper, an ethical framework is discussed, and the issues of bias in healthcare are investigated with great depth and detail.

1. "Ethics and Governance of Artificial Intelligence for Health" [95]
2. "Assessment List for Trustworthy Artificial Intelligence (ALTAI)" [96]
3. "NIST Trustworthy and Responsible AI" [97]
4. "Ethical Framework for Harnessing the Power of AI in Healthcare and Beyond" [98]

Key Requirements for "Trustworthy AI"

ALTAI:

- human agency and oversight
- technical robustness and safety
- privacy and data governance
- transparency
- diversity, non-discrimination and fairness
- environmental and societal well-being
- accountability

WHO:

- adopt regulations, standards, and best practices
- privacy by design and privacy by default
- confidentiality
- safety & risk assessments
- transparency
- bias
- data management
- infrastructure for AI applications and technical capacity
- evaluate and improve performance
- regular review

- intended use
- responsible and proficient use

- patient agency and perseverance of human authority
- ethical issues
- equal access
- assign liability

NIST- AI 100-2e2023:

- valid and reliable
- safe
- secure and resilient
- privacy-enhanced
- explainable and interpretable
- fair – harmful bias mitigated
- accountable and transparent

Ethical Framework for Harnessing the Power of AI in Healthcare and Beyond:

- Sensitivity:
  - Privacy
  - Accessibility
  - Inclusivity
- Evaluation:
  - Fairness
  - Non-Discriminative
  - Risk Assessment
- User Centric:
  - Contextual Intelligence
  - Emotional Intelligence
- Responsible:
  - Transparency
  - Accountability
  - Explainability
- Beneficence:
  - Sustainability
  - Resilience
  - Robustness
  - Reliability
- Security:
  - Adversarial Testing
  - Auditing

Consolidated List

- Beneficence and intended use (including responsible and proficient use, contextual intelligence)
- Human agency and perseverance of human authority and oversight
- Privacy, confidentiality
- Reliability, accountability, and liability
- Performance (including improvements, regular review, audit)
- Transparency (including explainability, interpretability)

- Diversity, fairness, accessibility (including ethical use)
- Sustainability
- Technical robustness, resilience and safety (including risk assessment, infrastructure capacity, data management and governance, adversarial testing, auditing)

Conclusions from the Healthcare Literature

There are different points of view and not all frameworks include all requirements. Interestingly, only the WHO publication mentions responsible and proficient use. None of these publications consider the human's responsibility to understand the model's intended use, like its strength, limitations and constraints, or even go into the topic of prompt guidelines to retrieve the best possible result from the system. Approaching an intelligent application with intuition might be more challenging than providing a manual (currently, disclaimers are replacing manuals). Again, only the WHO mentions liability, which is tightly intertwined with the AI application and its professional use. Note: only the WHO and the Ethical Framework for Harnessing the Power of AI in Healthcare and Beyond are addressing healthcare in particular. They reflect that AI applications in healthcare are expected to be validated with more scrutiny.

Sustainability is only once mentioned in the above discussed publications. This is surprising with the number of AI applications popping up throughout industries and regions and with sustainability being discussed in several frameworks. It reflects that the investment into AI in healthcare regarding performance is expected to exceed its (environmental) costs.

Bias in Healthcare

To avoid certain biases, some data must be depersonalized. While race, for example, can cause bias in a data set, it might also be a crucial information to provide safe and successful treatment. The handling of medical data is extremely complex and its purposeful evaluation is crucial. There are different aspects of bias, like data-driven, systematic, generalization, and human biases. The balance between the inclusion of characteristics and the avoidance of bias can be achieved by developing and adopting Explainable Artificial Intelligence (XAI) [2]. Techniques such as LIME[3] and SHAP[4] are examples of post-hoc explainability methods commonly used in healthcare. Moreover, it can also facilitate regulatory compliance by enabling audits and assessments of AI models [96].

Explainability seems to be a promising technique to mitigate bias in healthcare applications overall. Hence, the development of XAI is a major contribution to "Trustworthy AI" in healthcare.

[3] **LIME** (Local Interpretable Model-agnostic Explanations): LIME helps us understand why a Machine Learning model makes specific predictions. It does this by creating easy-to-understand explanations for individual predictions, even if the model itself is complex. Think of it as a way to peek into the model's decision-making process for each case.
[4] **SHAP** (SHapley Additive exPlanations): SHAP is another tool for explaining Machine Learning models. It tells us which features (like age, income, etc.) are most important in making predictions. It helps us see the big picture of how different factors influence the model's decisions, making it easier to understand and trust.

Further Applications of ML/AI in Healthcare

ML/AI applications can also be used to streamline the regulatory process [99], optimize the supply chain, assist in developing drugs and biological products [100] and improve direct patient care (improve medical treatment), indirect patient care (improve workflows in hospitals) and in-home care (wearable devices and sensors can assess and predict patient needs) [101]. There are regulations, standards, best practices and guidelines to develop medical devices with embedded ML/AI applications from several countries. ML/AI applications also can help improve manufacturing processes in the pharmaceutical industry [102].

Part III: AI Resilience Reframed: Benchmarking Model Inspired by Evolution

The goal of this part is to establish a novel framework to tackle the challenges and priorities in rating AI quality in order to future-proof AI systems. Evolution has done an unsurpassed job in selecting performance traits while maintaining the ability to survive. Exploring concepts in psychology reveals similarities between characteristics of materials, human behavior, and a possible new way to enhance resilience in AI technology. This chapter emphasizes the importance of policy makers, regulatory bodies, and governments overseeing AI development.

Let's start this part by looking at biological evolution compared to AI development with a focus on resilience, then shedding light on differences between human intelligence (HI) and artificial intelligence (AI), and closing the gap by looking at resilience from a psychological perspective. This part concludes with thoughts on implementing and measuring resilience in AI.

Comparison: Biological Evolution vs. AI Development

In biological evolution, new features (mutations) undergo testing for performance (adapted to a specific task) and resilience (persistence and advantage over time: survival). Organisms that persist over time exhibit a built-in protection against evolution. This might sound counterintuitive, but selection through different lenses (male/female or performance/resilience) makes a system more capable of maintaining its functional integrity from a holistic perspective [103].

Similarly, AI performance pertains to an AI's output in a predefined context, while AI resilience encompasses generalization (avoidance of overfitting) and adaptability to new tasks. While a market-driven industry is likely to neglect anything that doesn't drive revenue, it is the regulatory bodies' task to regulate and oversee the safety and hence, resilience of AI technology.

While AI applications undergo further evolution, more systems enabled for continuous learning after deployment, will conquer the market. The required degree of AI resilience of such a dynamic system exceeds that of a static system by far and wide.

AI resilience is a complex trait and might be neglected in the sheer awe of tempting performance. Hence, regulatory interventions become necessary to balance innovation and regulation.

Diversity and Resilience in AI Systems

Diversity is nature's answer to problem-solving. Therefore, it's of utmost importance that AI resilience is mandatory and regulated. Individual and unique approaches must be encouraged and rewarded. Only diverse AI technologies with solid but individual AI resilience solutions enhance global security.

> *"It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is most adaptable to change."*

This quote is misattributed to Darwin but still holds true, also in the perspective of AI systems; it is not performance that contributes to survival but, ultimately, AI resilience.

Augmenting intrinsic AI resilience with additional guardrails for the end-user, such as recommended use (a manual), adequate training, warnings, and (if possible) technical prevention from "off-label" usage is crucial and simple to provide – but frequently overlooked.

Policymakers, regulatory bodies, and governments must prioritize AI resilience in quality ratings to mitigate risks and ensure safe and future-proof AI integrations. Developing standardized metrics for evaluating resilience is essential.

The Challenge of Benchmarking AI Resilience

AI benchmarking approaches the saturation of traditional performance benchmarks ("fitness for intended use") [104], [105] with some systems already surpassing human baseline performance [106]. Stanford is leading the field with their Center for Research on Foundation Models using HELM [107], which evaluates models according to (currently) 87 scenarios and 50 metrics. The focus is on performance and the prevention of harm. Resilience is checked by evaluating how well a model performs when presented with two very different datasets (IMDB and BoolQ); the focus is on the ability to generalize while maintaining performance.

AI Resilience - Suggested Definition

We present a more holistic approach to AI resilience ultimately leading to suggesting an AI resilience score. In this context it is important to note that, in Psychology, there are intrinsic difficulties to measure resilience [108]. A useful definition from psychology is [109]: *"Resilience is the ability to resist stressors, to bounce back from stressors, and to grow from stressors."*

Please note that AI Resilience encompasses the ability to resist (resistance), the ability to bounce back (resilience) and to grow from stressors (plasticity):

Resistance to a stressor can be likened to the "stiffness" of a material but also to the diversified and highly dynamic approach of the human immune system. Hence, resistance has two contradicting aspects, both having their rightful usefulness. Survival is not the absence of challenges but the (shared) responsibility to

face them proactively and sustainably.

Resilience is the process of bouncing back from the impact of a stressor over time, influenced by factors like the magnitude and duration of the stressful event (external factors) and the elasticity/adaptability of the stressed subject (internal factors). Resilience is dynamic and affected by various variables. However, there are instances where the impact of a stressor exceeds the ability to restore original functionality.

Plasticity refers to a permanent change. It can be dysfunctional, like trauma in a psychological context, a fracture (of a bone) in medicine or a point of failure in material sciences. Or, it can be functional, such as in showing increased performance/resilience due to training [87].

The following definition of resilience for AI technology is suggested.

AI resilience consists of a system's resistance, resilience, and plasticity.

AI resistance reflects the system's ability to maintain a required minimal performance in the face of intrusion, manipulation, misuse, and abuse.

AI resilience focuses on the time, capacity, and capability needed to bounce back to the required minimal performance after an incident.

AI plasticity serves as the system's gauge indicating its tolerance to "make it or break it" and allows quick action in the case of system failure or allows continuously improving AI resilience.

Unsurprisingly, misuse, abuse, and incidents with AI applications soar! The AI use cases exemplify that incidents do occur despite the awareness to manage and mitigate risks.

However, integrating AI into a Quality Management System to control, improve, correct and prevent actions and/or risks is challenging as the criteria upon which the AI is judged are unclear. In regulated industries, a third-party rating would enhance safety beyond the validation of the AI, which is currently done under the aspect of "fit for use."

Proposed AI Resilience Score

A resilience score from 0 to 10 is suggested that reflects an AI's resilience considering its three pillars: resistance, resilience, and plasticity. Such a score could look like (for example) 16:5-8-3 representing the sum of the three pillars and each of the three pillars, separately. The distribution of the scores of the three pillars could reflect the diversity of different AI systems. This would allow a more informed decision regarding risks and their mitigation if/when combining different AI systems.

The focus of policymakers, risk managers, regulatory bodies, and governments must prioritize AI resilience over the aspect of performance and reward any step taken in this direction, promoting diverse solutions to increase the diversity of AIs.

Let's shift the focus now to the interaction between AI and humans.

Intelligence Awareness

The concept of "Intelligence Awareness" [110] emphasizes understanding differences in intelligence rather than comparing them. Intelligence Awareness is not a widely used or known concept (yet) and is clearly distinct from the concept of Harvard psychologist Howard Gardner who introduced different aspects of intelligence [111], [112]. “Intelligence Awareness” highlights the need for humans to learn to interact with other intelligent systems safely and efficiently, by respecting each other’s different abilities. A beautiful example is the science fiction book “Project Hail Mary” by the bestselling author Andy Weir [113]. As AI approaches or surpasses human performance, its benchmarking becomes crucial. There is diversity in intelligent systems, and respect for each specific ability enhances safety and effectiveness. In the next section, the fundamental differences are explored.

Fundamental Differences in Intelligent Systems

Comparing artificial intelligence (AI) and human intelligence (HI) [114] assumes they can be compared, with HI currently seen as the gold standard. However, the biological basis of HI differs significantly from AI's highly precise silicon chip foundation. This difference in the hardware impacts the fundamental functionality of the two forms of intelligence. A leap will be observed once AI can be trained and run on quantum computers or biological computers [115], as both approaches will merge the silicon chip characteristics with the quantum abilities of the human brain [116], [117], [118], [119]. Such AI systems will likely have the performance of current AI combined with the ability of the human brain to solve complex tasks. It is noteworthy that both approaches aim at combining deterministic computing with non-deterministic approaches. At that point, the question arises: how do we judge an intelligence that can generate answers that humans might not even be able to understand anymore, like the famous “42” from The Hitchhiker’s Guide Through the Galaxy [120].

Bibliography

[1] M. W. Dictionary, "Merriam Webster Dictionary," [Online]. Available: https://www.merriam-webster.com/dictionary/governance. [Accessed 24 02 2024].

[2] C. Dictionary, "Cambridge Dictionary," [Online]. Available: https://dictionary.cambridge.org/dictionary/english/compliance. [Accessed 24 02 2024].

[3] IBM, "What is explainable AI?," IBM, [Online]. Available: https://www.ibm.com/topics/explainable-ai?utm_content=SRCWW&p1=Search&p4=43700074359379082&p5=e&gclid=CjwKCAjw4ZWkBhA4EiwAVJXwqaOswoxlekelxe20HE0gNhPjlU09SzOtlJ888FRz91kTGBO2tRsZZBoC_aAQAvD_BwE&gclsrc=aw.ds. [Accessed 14 04 2024].

[4] P. S. M. M. Prashant Gohel, "Explainable AI: current status and future directions," 12 07 2021. [Online]. Available: https://arxiv.org/abs/2107.07045. [Accessed 14 04 2024].

[5] M. a. W. S. a. Z. A. a. B. P. a. V. L. a. H. B. a. S. E. a. R. I. D. a. G. T. Mitchell, "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 2019, p. 220–229.

[6] E. O. O. T. PRESIDENT, "MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES - Advancing Governance, Innovation, and Risk Management for Agency Use of," The Director, Washington, D. C., 2024.

[7] The University of Queensland, Australia, "History of Artificial Intelligence," The University of Queensland, Australia, [Online]. Available: https://qbi.uq.edu.au/brain/intelligent-machines/history-artificial-intelligence. [Accessed 14 04 2024].

[8] IBM, "What is machine learning?," [Online]. Available: https://www.ibm.com/topics/machine-learning. [Accessed 24 02 2024].

[9] tinyML Foundation, "tinyML Foundation," [Online]. Available: https://www.tinyml.org/about/. [Accessed 24 02 2024].

[10] IBM, "What is generative AI?," [Online]. Available: https://research.ibm.com/blog/what-is-generative-AI. [Accessed 24 02 2024].

[11] IBM, "What is strong AI?," [Online]. Available: https://www.ibm.com/topics/strong-ai. [Accessed 24 02 2024].

[12] proft.me, "Types of machine learning algorithms," [Online]. Available: https://en.proft.me/2015/12/24/types-machine-learning-algorithms/. [Accessed 02 03 2024].

[13] C. W. Xiang D, "Privacy Protection and Secondary Use of Health Data: Strategies and Methods," *Biomed Res Int.,* 07 10 2021.

[14] Wikipedia, "Wisdom of the crowd," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Wisdom_of_the_crowd. [Accessed 24 02 2024].

[15] X, "X Terms of Service," X, 29 09 2023. [Online]. Available: https://twitter.com/en/tos. [Accessed 02 03 2024].

[16] D. H. a. t. a. press, "Reddit has struck a $60m deal with Google that lets the search giant train AI models on its posts," Fortune, 23 02 2024. [Online]. Available: https://fortune.com/2024/02/23/reddit-60m-deal-google-search-giant-train-ai-models-on-posts/. [Accessed 02 03 2024].

[17] K. Coar, "open source initiative," 08 02 2004. [Online]. Available: https://opensource.org/license/apache-2-0. [Accessed 02 03 2024].

[18] open source initiative, "The MIT License," open source initiative, [Online]. Available: https://opensource.org/license/mit. [Accessed 02 03 2024].

[19] Europäisches Patentamt, "Artificial intelligence and machine learning," Europäisches Patentamt, [Online]. Available: https://www.epo.org/en/legal/guidelines-epc/2023/g_ii_3_3_1.html. [Accessed 02 03 2024].

[20] The PatentLawyer, "EPO updates guidelines for examining AI inventions," 20 02 2024. [Online]. Available: https://patentlawyermagazine.com/epo-updates-guidelines-for-examining-ai-inventions/. [Accessed 14 04 2024].

[21] I. Guttmann, "METHOD AND SYSTEM TO SAFELY GUIDE INTERVENTIONS IN PROCEDURES THE SUBSTRATE WHEREOF IS NEURONAL PLASTICITY". Europe 01 04 2022.

[22] S. W. &. J. Grasser, "Japan's New Draft Guidelines on AI and Copyright: Is It Really OK to Train AI Using Pirated Materials?," SQUIRE, 12 03 2024. [Online]. Available: https://www.privacyworld.blog/2024/03/japans-new-draft-guidelines-on-ai-and-copyright-is-it-really-ok-to-train-ai-using-pirated-materials/. [Accessed 01 04 2024].

[23] P. D. T. (. Law), "Generative AI and Copyright Infringement," NUS - National University of Singapore, 01 2024. [Online]. Available: https://law.nus.edu.sg/trail/generative-ai-copyright-infringement/. [Accessed 14 04 2024].

[24] J. Vincent, "Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day," The Verge, 24 03 2026. [Online]. Available: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist. [Accessed 03 03 2024].

[25] P. Lee, "Learning from Tay's introduction," Microsoft, 25 03 2016. [Online]. Available: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/. [Accessed 03 03 2024].

[26] Wikipedia, "Tay (chatbot)," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Tay_(chatbot). [Accessed 03 03 2024].

[27] J. Dastin, "https://globalnews.ca/news/4532172/amazon-jobs-ai-bias/," Global News, 10 10 2018. [Online]. Available: https://globalnews.ca/news/4532172/amazon-jobs-ai-bias/. [Accessed 02 03 2024].

[28] J. Vincent, "Amazon reportedly scraps internal AI recruiting tool that was biased against women," The Verge, 10 10 2018. [Online]. Available: https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report. [Accessed 02 03 2024].

[29] S. W. a. H. Schellmann, "LinkedIn's job-matching AI was biased. The company's solution? More AI.," MIT Technology Review, 23 06 2021. [Online]. Available: https://www.technologyreview.com/2021/06/23/1026825/linkedin-ai-bias-ziprecruiter-monster-artificial-intelligence/. [Accessed 14 04 2024].

[30] R. L. I. P. F. S. a. I. U. Trisha Thadani, "The final 11 seconds of a fatal Tesla Autopilot crash: A reconstruction of the wreck shows how human error and emerging technology can collide with deadly results," The Washington Post, 06 10 2023. [Online]. Available: https://www.washingtonpost.com/technology/interactive/2023/tesla-autopilot-crash-analysis/. [Accessed 04 03 2024].

[31] B. P. C. V. a. S. M. Ziad Obermeyer, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science,* vol. 366, no. 6464, pp. 447-453, 25 10 2019.

[32] T. Telford, "Apple Card algorithm sparks gender bias allegations against Goldman Sachs," The Washington Post, 11 11 2019. [Online]. Available: https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/. [Accessed 02 03 2024].

[33] BBC, "Apple's 'sexist' credit card investigated by US regulator," BBC, 11 11 2019. [Online]. Available: https://www.bbc.com/news/business-50365609. [Accessed 24 02 2024].

[34] WIRED, "The Apple Card Didn't 'See' Gender—and That's the Problem," WIRED, 19 11 2019. [Online]. Available: https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/. [Accessed 24 02 2024].

[35] N. Vigdor, "Apple Card Investigated After Gender Discrimination Complaints," The New York Times, 10 11 2019. [Online]. Available: https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html. [Accessed 24 02 2024].

[36] J. Vincent, "Apple's credit card is being investigated for discriminating against women," The Verge, 11 11 2019. [Online]. Available: https://www.theverge.com/2019/11/11/20958953/apple-credit-card-gender-discrimination-algorithms-black-box-investigation. [Accessed 24 02 2024].

[37] New York State Department of Financial Services, "Report on Apple Card Investigation," New York State Department of Financial Services, 2021.

[38] I. C. Campbell, "The Apple Card doesn't actually discriminate against women, investigators say," The Verge, 24 03 2021. [Online]. Available: https://www.theverge.com/2021/3/23/22347127/goldman-sachs-apple-card-no-gender-discrimination. [Accessed 02 03 2024].

[39] W. D. Heaven, "Predictive policing algorithms are racist. They need to be dismantled.," MIT Technology Review, 17 07 2020. [Online]. Available: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/. [Accessed 02 03 2024].

[40] A. Zilber, "Air Canada ordered to refund passenger after 'misleading' conversation with site's AI chatbot," New York Post, 19 02 2024. [Online]. Available: https://nypost.com/2024/02/19/business/air-canada-ordered-to-refund-passenger-after-ai-chatbots-misleading-messages/. [Accessed 03 03 2024].

[41] A. Belanger, "Air Canada must honor refund policy invented by airline's chatbot," arsTECHNICA, 16 02 2024. [Online]. Available: https://arstechnica.com/tech-policy/2024/02/air-canada-must-honor-refund-policy-invented-by-airlines-chatbot/. [Accessed 02 03 2024].

[42] E. Napolitano, "UnitedHealth uses faulty AI to deny elderly patients medically necessary coverage, lawsuit claims," MONEYWATCH, 20 11 2023. [Online]. Available: https://www.cbsnews.com/news/unitedhealth-lawsuit-ai-deny-claims-medicare-advantage-health-insurance-denials/. [Accessed 02 03 2024].

[43] B. Pierson, "Lawsuit claims UnitedHealth AI wrongfully denies elderly extended care," Reuters, 14 11 2023. [Online]. Available: https://www.reuters.com/legal/lawsuit-claims-unitedhealth-ai-wrongfully-denies-elderly-extended-care-2023-11-14/. [Accessed 02 03 2024].

[44] S. P. a. D. Hassabis, "Our next-generation model: Gemini 1.5," Google, 15 02 2024. [Online]. Available: https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#gemini-15. [Accessed 14 04 2024].

[45] J. L. S. G. a. R. M. Davey Alba, "Google Left in 'Terrible Bind' by Pulling AI Feature After Right-Wing Backlash," TIME, 28 02 2024. [Online]. Available: https://time.com/6835975/google-gemini-backlash-bias/. [Accessed 02 03 2024].

[46] SAE Blog, "SAE Levels of Driving Automation™ Refined for Clarity and International Audience," SAE, 03 05 2021. [Online]. Available: https://www.sae.org/blog/sae-j3016-update. [Accessed 24 02 2024].

[47] EUR-Lex, "Regulation - 2019/2144 - EN - EUR-Lex," EUR-Lex, 05 09 2022. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2019/2144/oj#d1e1549-1-1. [Accessed 24 02 2024].

[48] ISO, "ISO/CD PAS 8800: Road Vehicles - Safety and artificial intelligence," ISO, [Online]. Available: https://www.iso.org/standard/83303.html. [Accessed 24 02 2024].

[49] Fraunhofer Institute for Cognitive Systems IKS, "AI regulation and AI standardization," Fraunhofer Institute, [Online]. Available: https://www.iks.fraunhofer.de/en/topics/artificial-intelligence/ai-standardization.html. [Accessed 24 02 2024].

[50] ISO, "ISO/IEC TR 5469:2024," ISO, 01 2024. [Online]. Available: https://www.iso.org/standard/81283.html. [Accessed 24 02 2024].

[51] ISO, "ISO/IEC AWI TS 22440," ISO, [Online]. Available: https://www.iso.org/standard/87118.html. [Accessed 24 02 2024].

[52] I. E. Team, "New standard to increase safety of AI," International Electrotechnical Commisson, 16 01 2024. [Online]. Available: https://www.iec.ch/blog/new-standard-increase-safety-ai. [Accessed 24 02 2024].

[53] ISO, "ISO/TR 4804:2020," ISO, 12 2020. [Online]. Available: https://www.iso.org/standard/80363.html. [Accessed 24 02 2024].

[54] ISO, "ISO/IEC 27000:2018," ISO, 2018. [Online]. Available: https://www.iso.org/standard/73906.html. [Accessed 02 03 2024].

[55] ISO, "ISO/IEC 42001:2023," ISO, 2023. [Online]. Available: https://www.iso.org/standard/81230.html. [Accessed 14 04 2024].

[56] NIST, "Artificial Intelligence Risk Management," NIST, 01 2023. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. [Accessed 14 04 2024].

[57] UK Civil Aviation Authority, "The CAA's strategy for Artificial Intelligence (AI)," CAA, [Online]. Available: https://www.caa.co.uk/our-work/innovation/artificial-intelligence/. [Accessed 14 04 2024].

[58] T. T. Pham, "Chief Scientist and Technical Advisor for Artificial Intelligence - Machine Learning," [Online]. Available: https://www.faa.gov/aircraft/air_cert/step/disciplines/pham_bio. [Accessed 14 04 2024].

[59] H. Weitering, "Beyond Automation: How AI Is Transforming Aviation," 14 06 2023. [Online]. Available: https://www.ainonline.com/aviation-news/aerospace/2023-06-14/beyond-automation-how-ai-transforming-aviation. [Accessed 14 04 2024].

[60] European Union, "CORDIS results pack on AI in science," [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/c0e52bea-5bb0-11ee-9220-01aa75ed71a1. [Accessed 14 04 2024].

[61] ISA - International Society of Automation, "SA/IEC 62443 Series of Standards: The World's Only Consensus-Based Automation and Control Systems Cybersecurity Standards," ISA - International Society of Automation, [Online]. Available: https://www.isa.org/standards-and-publications/isa-standards/isa-iec-62443-series-of-standards. [Accessed 24 02 2024].

[62] IEC - International Electrotechnical Commission, "IEC TS 62351-100-4:2023," International Electrotechnical Commission, 2023. [Online]. Available: https://webstore.iec.ch/publication/63323. [Accessed 24 02 2024].

[63] IEC - International Electrotechnical Commission, "IEC TR 61850-90-4:2020," International Electrotechnical Commission, 2020. [Online]. Available: https://webstore.iec.ch/publication/64801. [Accessed 24 02 2024].

[64] enisa, "Cybersecurity and privacy in AI - Forecasting demand on electricity grids," enisa, 07 06 2023. [Online]. Available: https://www.enisa.europa.eu/publications/cybersecurity-and-privacy-in-ai-forecasting-demand-on-electricity-grids. [Accessed 24 02 2024].

[65] M. O. Y. R. S. M. N. K. S. Z. W. W.-Y. M. Feras A. Batarseh, "Realtime Management of Wastewater Treatment Plants Using AI," Virginia Tech & DC Water, 2022. [Online]. Available: https://www.waterrf.org/sites/default/files/file/2022-11/2022_IWS-Challenge-Solution_Virginia-Tech.pdf. [Accessed 24 02 2024].

[66] P. C. o. A. o. S. &. Technology, "Strategy for Cyber-Physical Resilience: Fortifying Our Critical Infrastructure for a Digital World," Executive Office of the President, 2024.

[67] The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," The White House, 30 10 2023. [Online]. Available: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. [Accessed 24 02 2024].

[68] America's Cyber Defense Agency, "Artificial Intelligence," America's Cyber Defense Agency, [Online]. Available: https://www.cisa.gov/ai. [Accessed 24 02 2024].

[69] European Commission, "Artificial Intelligence Act," European Commission, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206. [Accessed 24 02 2024].

[70] European Commission, "Annexes to the EU AI Act," European Commission, 2021. [Online]. Available: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_2&format=PDF. [Accessed 2024 02 2024].

[71] European Parliament, "Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI," European Parliament, 02 12 2023. [Online]. Available: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai. [Accessed 24 02 2024].

[72] OECD, "Accountability (Principle 1.5)," OECD.AI Policy Observatory, [Online]. Available: https://oecd.ai/en/dashboards/ai-principles/P9. [Accessed 24 02 2024].

[73] Government of Canada, "The Artificial Intelligence and Data Act (AIDA)," 09 2023. [Online]. Available: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document. [Accessed 14 04 2024].

[74] J. Hoppe, "The Dropping of the TURDSID in Vietnam," US Naval Institute, 10 2021. [Online]. Available: https://www.usni.org/magazines/naval-history-magazine/2021/october/dropping-turdsid-vietnam. [Accessed 02 03 2024].

[75] U.S. Department of Defense , "New Strategy Outlines Path Forward for Artificial Intelligence," U.S. Department of Defense , 12 02 2019. [Online]. Available: https://www.defense.gov/News/Releases/Release/Article/1755388/new-strategy-outlines-path-forward-for-artificial-intelligence/. [Accessed 02 03 2024].

[76] R. Gigova, "Who Vladimir Putin thinks will rule the world," CNN, 02 09 2017. [Online]. Available: https://www.cnn.com/2017/09/01/world/putin-artificial-intelligence-will-rule-world/index.html. [Accessed 02 03 2024].

[77] Congressional Research Service (CRS), "Artificial Intelligence and National Security," 2018.

[78] E. A. A. &. C. R. Baiz, "Generative AI in Education and Research: Opportunities, Concerns, and Solutions," *J. Chem. Educ.,* vol. 100, no. 8, p. 2965–2971, 27 07 2023.

[79] K. A. B. D. G.-R. A. R. M. Bozkurt A, "Artificial Intelligence and Reflections from Educational Landscape: A Review of AI Studies in Half a Century," *Sustainability,* vol. 13(2), no. 800, 2021.

[80] W. Kenton, "Lehman Brothers: History, Collapse, Role in the Great Recession," Investopedia, 31 12 2022. [Online]. Available: https://www.investopedia.com/terms/l/lehman-brothers.asp. [Accessed 24 02 2014].

[81] A. R. Sorkin, Too Big to Fail: Inside the Battle to Save Wall Street, Penguin, 2010.

[82] Congressional Research Service (CRS), "Systemic Risk And The Long-Term Capital Management Rescue," Congressional Research Service, 1999.
[83] BIS, "Studies on the Validation of Internal Rating Systems," BIS - Bank for International Settlements, 2005.
[84] BIS, "Studies on the Validation of Internal Rating Systems (revised)," BIS - Bank for International Settlements, 2005.
[85] Board of Governors of the Federal Reserve System, "Supervision and Regulation Letters - SR 11-7: Guidance on Model Risk Management," 04 04 2011. [Online]. Available: https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm. [Accessed 24 02 2024].
[86] M. Heusser, "Software Testing Lessons Learned From Knight Capital Fiasco," CIO, 14 08 2012. [Online]. Available: https://www.cio.com/article/286790/software-testing-lessons-learned-from-knight-capital-fiasco.html. [Accessed 24 02 2024].
[87] govtrack.us, "H.R. 15073 (91st): An Act to amend the Federal Deposit Insurance Act to require insured banks to maintain certain records, to require that certain transactions in U.S. currency be reported to the Department of the Treasury, and for other purposes," 26 11 1970. [Online]. Available: https://www.govtrack.us/congress/bills/91/hr15073/text. [Accessed 14 04 2024].
[88] congress.gov, "H.R.3162 - Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (USA PATRIOT ACT) Act of 2001," 24 10 2001. [Online]. Available: https://www.congress.gov/bill/107th-congress/house-bill/3162. [Accessed 14 04 2024].
[89] ECB, "ECB updates Guide to internal models," 19 02 2024. [Online]. Available: https://www.bankingsupervision.europa.eu/press/pr/date/2024/html/ssm.pr240219~8c10a7d827.en.html. [Accessed 14 04 2024].
[90] McKinsey & Company, "Model Risk Management," 2019.
[91] Security Standards Council, "PCI DSS," Security Standards Council, [Online]. Available: https://www.pcisecuritystandards.org/document_library/?document=pci_dss. [Accessed 02 03 2024].
[92] Security Standards Council, "PCI 3DS," Security Standards Council, [Online]. Available: https://www.pcisecuritystandards.org/document_library/?document=3DS_standard. [Accessed 02 03 2024].
[93] Microsoft, "Microsoft Azure Compliance Offerings," [Online]. Available: Azure - Compliance Offerings. [Accessed 14 04 2024].
[94] W. Daniel, "https://fortune.com/2023/05/26/jpmorgan-indexgpt-a-i-stock-picker/," FORTUNE, 26 05 2023. [Online]. Available: https://fortune.com/2023/05/26/jpmorgan-indexgpt-a-i-stock-picker/. [Accessed 02 03 2024].
[95] WHO guidance, "Ethics and Governance of Artificial Intelligence for Health," WHO, 2021.
[96] European Commission, "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment," European Commission, 2020.
[97] NIST, "NIST Trustworthy and Responsible AI - NIST AI 100-2e2023," NIST, 2024.
[98] S. &. K. R. &. B. S. Nasir, "Ethical Framework for Harnessing the Power of AI in Healthcare and Beyond," 08 2023. [Online]. Available: https://www.researchgate.net/publication/373641885_Ethical_Framework_for_Harnessing_the_Power_of_AI_in_Healthcare_and_Beyond. [Accessed 24 02 2024].
[99] A. Bilea, "How AI is Revolutionizing Pharma Regulatory Compliance," LinkedIn, 26 07 2023. [Online]. Available: https://www.linkedin.com/pulse/how-ai-revolutionizing-pharma-regulatory-compliance-anca-bilea/. [Accessed 24 02 2024].

[100] FDA, "Using Artificial Intelligence & Machine Learning in the Development of Drug & Biological Products," FDA.

[101] CAPRA - Canadian Association of Professionals in Regulatory Affairs, "Artificial Intelligence – Revolutionizing the Healthcare Industry," CAPRA - Canadian Association of Professionals in Regulatory Affairs, 27 10 2023. [Online]. Available: https://capra.ca/en/blog/artificial-intelligence-revolutionizing-the-healthcare-industry-2023-10-27. [Accessed 24 02 2024].

[102] FDA, "Artificial Intelligence in Drug Manufacturing," FDA, [Online]. Available: https://www.fda.gov/media/165743/download. [Accessed 24 02 2024].

[103] V. P. Shcherbakov, "Biological species is the only possible form of existence for higher organisms: the evolutionary meaning of sexual reproduction," *Biol Direct.,* vol. 5, no. 14, 22 03 2010.

[104] Stanford University - Human-Centered Artificial Intelligence, "The AI Index Report - Measuring trends in Artificial Intelligence," Stanford University - Human-Centered Artificial Intelligence, 2023. [Online]. Available: https://aiindex.stanford.edu/report/. [Accessed 24 02 2024].

[105] aqua, "AI Benchmark Ranking – The Ultimate Guide to Comparing and Evaluating AI Performance," Aquarius, 01 12 2023. [Online]. Available: https://aquariusai.ca/blog/ai-benchmark-ranking-the-ultimate-guide-to-comparing-and-evaluating-ai-performance. [Accessed 24 02 2024].

[106] S. Lynch, "AI Benchmarks Hit Saturation," Standford University. HAI - Human-Centered Artificial Intelligence, 03 04 2023. [Online]. Available: https://hai.stanford.edu/news/ai-benchmarks-hit-saturation. [Accessed 24 02 2024].

[107] Center for research on Foundation Models, "HELM," Stanford University, [Online]. Available: https://crfm.stanford.edu/helm/lite/latest/. [Accessed 02 03 2024].

[108] B. K. N. J. Windle G, "A methodological review of resilience measurement scales," *Health Qual Life Outcomes,* vol. 9, no. 8, 04 02 2011.

[109] Y. H. Ruud J.R. Den Hartigh, "Conceptualizing and measuring psychological resilience: What can we learn from physics?," *New Ideas in Psychology,* vol. 66, no. 100934, 2022.

[110] G. C. v. d. B.-V. R. A. M. B. e. a. J. E. (Hans). Korteling, "Human versus Artificial Intelligence," *Front. Artif. Intell.,* vol. 4, 25 03 2021.

[111] K. Cherry, "Gardner's Theory of Multiple Intelligences," 11 03 2023. [Online]. Available: https://www.verywellmind.com/gardners-theory-of-multiple-intelligences-2795161. [Accessed 14 04 2024].

[112] Wikipedia, "Howard Gardner," [Online]. Available: https://en.wikipedia.org/wiki/Howard_Gardner#cite_note-Gordon,_Lynn_Melby_2006-1. [Accessed 14 04 2024].

[113] Wikipedia, "Project Hail Mary," [Online]. Available: https://en.wikipedia.org/wiki/Project_Hail_Mary. [Accessed 14 04 2024].

[114] V. Acharya, "AI vs. HI: The Battle of Intelligences – Exploring Advantages and Limitations," Medium, 24 07 2023. [Online]. Available: https://medium.com/@vishwasacharya/ai-vs-hi-the-battle-of-intelligences-exploring-advantages-and-limitations-89759bee090f. [Accessed 24 02 2024].

[115] A. Tongen, "Will Biological Computers Enable Artificially Intelligent Machines to Become Persons?," *Dignity,* vol. 9, no. 4, 2003.

[116] R. L. M.D., "Psychology Today," 02 08 2021. [Online]. Available: https://www.psychologytoday.com/ca/blog/biocentrism/202108/quantum-effects-in-the-brain. [Accessed 24 02 2024].

[117] Neuroscience News, "Our Brains Use Quantum Computation," Neuroscience News, 22 20 2022. [Online]. Available: https://neurosciencenews.com/brain-quantum-computing-21695/. [Accessed 24 02 2024].

[118] C. H. K. Koch, "Quantum mechanics in the brain," *Nature* , vol. 440, no. 611, 2006.

[119] Trinity College Dublin, "New research suggests our brains use quantum computation," Phys Org, 19 20 2022. [Online]. Available: https://phys.org/news/2022-10-brains-quantum.html. [Accessed 24 02 2024].

[120] Wikipedia, "The Hitchhiker's Guide to the Galaxy," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/The_Hitchhiker%27s_Guide_to_the_Galaxy. [Accessed 24 02 2024].