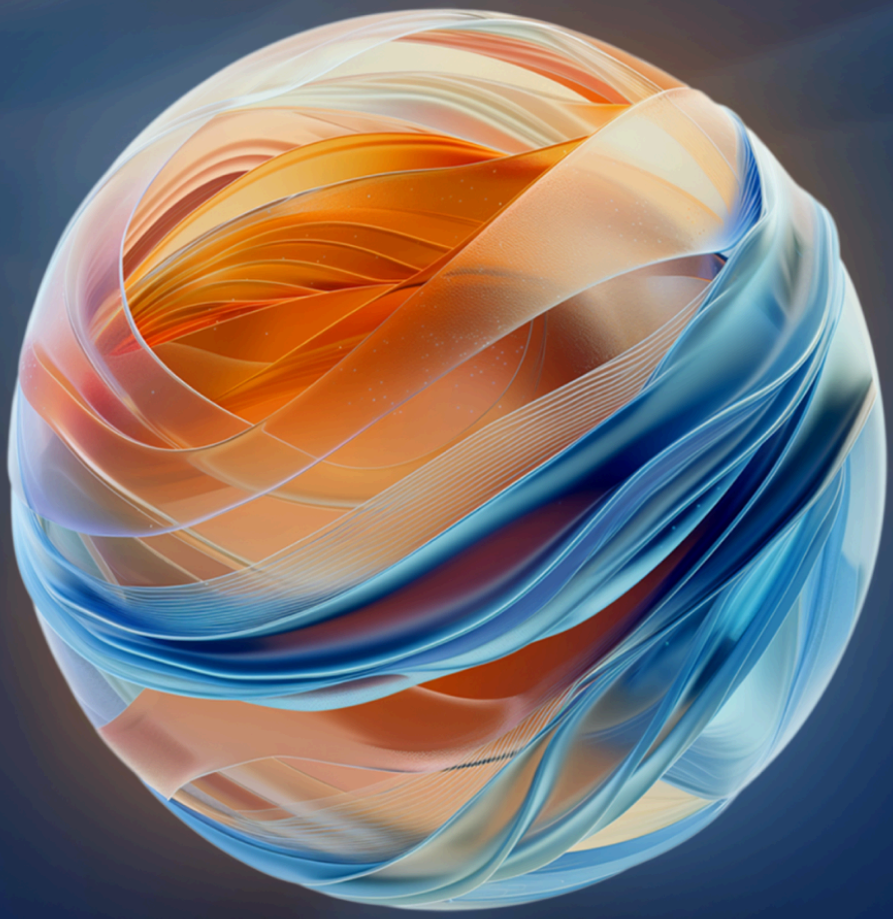
AI Risk Management:

Thinking Beyond Regulatory Boundaries



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 Author

Dr. Chantal Spleiss

Contributors

Candy Alexander
Sanitra Angram
Renu Bedi
Madhav Chablani
Filip Chyla
Dr. Marco Ermini
Becky Gaylord
Frederick Haenig
Arpitha Kaushik
Kathie Miley
Ashish Vashishtha
Peter Ventura

Reviewers

Harsh Daiya
Beniamino Di Martino
Debrup Ghosh
Rahul Gupta
Ashish Gupta
Dheeraj Gurugubelli
J. Winston Hayden
Sybil VR Kleinmichel
Vaibhav Malik
Krishna Manghat
Taresh Mehra
Sven Olensky
Meghana Parwate
Hena Prasanna
Chandra Rajagopalan
Michael Roza
Maria (MJ) Schwenger
Rakesh Sharma

CSA Global Staff

Ryan Gifford
Claire Lehnert
Stephen Lumpe

Table of Contents

AI Risk Management: Thinking Beyond Regulatory Boundaries

Introduction

In a world where artificial intelligence (AI) is becoming increasingly integrated into industry processes, the need for accurate, purposeful, and results-based AI auditing is mission-critical. While AI offers tremendous benefits, such as increased efficiency, augmented decision-making, and innovative capabilities, it also introduces significant risks and challenges for organizations of any size. Establishing trust in AI can only be achieved through a comprehensive approach to AI auditing that addresses compliance proactively with improvements beyond the regulatory necessities.

This paper presents a holistic overview and applicable methodology needed for impartially assessing intelligent systems designed to be applicable across industries. These audit considerations build upon existing AI audit best practices and provide an innovative approach as it spans the entire AI lifecycle and includes sample audit questions in the Appendix.

Privacy, security (including information and cybersecurity), and trustworthiness are emphasized by proposing a risk-based approach with a focus on critical and investigative thinking, curiosity, and the auditor's ability to assess systems for unintended behavior. The approach defined herein is aligned with audit best practices and is independent of current or future AI regulations but relevant to AI management systems. Taking the dynamic nature of AI evolution into consideration, the content of this paper outlines the areas an auditor needs to be aware of but refrains from going into details that are subject to change.

Novel audit methodologies [1], [2], [3] are currently surfacing, for example, to assess Large Language Models (LLMs) [4], [5], [6], and there is mutual consent for the need for innovative risk assessment. Our sister Working Group discusses the topic of AI governance from a fundamental perspective: Don't Panic! Getting Real About AI Governance [7]. This paper presents objectives that can be used to assess diverse intelligent systems along the whole lifecycle [8], [9] with or without embodiment. Further, these audit considerations facilitate a structured approach to an audit of intelligent systems by providing a thematic overview followed by questions in the corresponding appendices that help to promote outside-the-box thinking and conduct assessments beyond the scope of compliance.

Executive Summary

This paper has the ambition to contribute to the creation and evolution of AI auditing, providing considerations for a holistic approach to assess diverse, intelligent systems along their entire lifecycle and with or without embodiment. A methodology is presented to ensure intelligent systems are assessed rigorously, enabling companies to mitigate any detected risks and approach compliance innovatively and proactively to build trustworthy AI.

These AI audit considerations propose recommendations assisting in the assessment of intelligent systems across industries and legislations by emphasizing critical and investigative thinking and a curious yet analytic attitude of the auditor. Based on rigorous assessments, the benefits of introducing intelligent systems can be monetarily fully harvested with less uncertainty about hidden pitfalls leading to business impacts in the future.

The dynamic nature of AI is considered by giving an overview instead of a deep dive into each topic. This allows for change but also requires the auditor to keep specific knowledge updated or augment the auditor team with corresponding specialists. Each audit is as unique as its underlying business case; these considerations aim to present a methodology that can be used for diverse AI technology.

This paper first introduces considerations, giving an auditor an overview. In the Appendix, questions are proposed to address each topic during an audit or assessment, facilitating risk assessment beyond compliance and promoting outside-the-box thinking.

The primary audiences include senior management, regulators, and those who perform assessments and/or audits of intelligent systems.

Parts 1 and 2: How They Play Together

The first part of this proposal to assess intelligent systems explains fundamental concepts and establishes a comprehensive understanding of the components necessary and/or used. This foundational knowledge equips auditors with the basic knowledge to assess the trustworthiness of intelligent systems, covering a broad range of technologies as concisely as possible. This preparation enables critical thinking and facilitates risk assessment, exceeding the scope of typical compliance.

The second part comprises appendices with potential questions corresponding to each chapter in the first section of this paper. The practical lists with possible questions are not exhaustive but serve as a guideline to identify potential risks from a current perspective. By asking these questions, the aim is to stimulate unconventional thinking and challenge existing assumptions, thereby enhancing risk assessment practices and increasing overall trustworthiness in intelligent systems.

Definitions

The objective of this section is to provide the fundamental concepts, principles, and vocabulary used in this paper to assess AI end-to-end. This clarification shall support users and can calibrate the auditors' interpretation of these basic concepts and terms to support comparable assessment results in various countries and industries. The authors' intent is to help auditors effectively and efficiently assess the trustworthiness of AI.

The terms defined herein are included within topical areas, as referred to in this document, or within related AI assessment methodologies and are relevant to the purpose of this method.

The goal of any AI assessment is to reveal possible risks. The mitigation of these risks must be managed by the auditee and by addressing possible issues as the deployed AI matures in its trustworthiness. An important cornerstone towards the improvement of intelligent systems is AI Resilience.

AI Resilience: Robustness, Resilience, and Plasticity

AI resilience refers to an intelligent system's ability to maintain and recover its performance in the face of various threats or actual impacting events. AI resilience encompasses three main pillars [Figure 1]:

1. **Robustness**: The ability to withstand threats without compromising intended performance or functionality.
2. **Resilience**: The capability of a strained AI system to recover normal operations after a performance-impacting incident. This is the ability of an AI system to recover and return to its previous operational state (snap back).
3. **Plasticity**: The ability to recover from impacting incidents with functional changes or evolutionary adaptation. This can happen supervised or unsupervised; performance and resilience may increase the system's capability to meet novel challenges. It is also the system's gauge indicating its tolerance to "make it or break it."

While plasticity is becoming an interesting and powerful option, the changes must be traceable, and their functionality must be validated. If an AI becomes dysfunctional after an incident, with or without any self-repair options, it must be possible to switch it off (kill switch).

Together, these core pillars of AI resilience work together to ensure that an intelligent system can not only endure disruptions but also learn and evolve from them, improving the system over time.

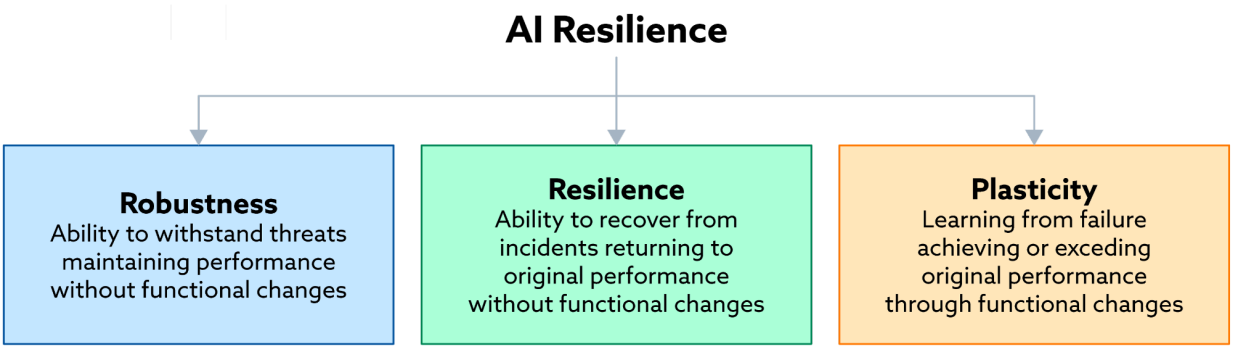


Figure 1: AI Resilience [10], [11]

Intelligent Systems: AI Systems, AI Agents, and Robots

AI System: A broad term encompassing software-based applications that utilize AI to perform tasks such as data analysis, decision-making, and natural language processing. These systems can operate without physical embodiment and are often found in areas like finance, healthcare, and customer service.

AI Agent: A more specific type of intelligent system designed to autonomously perceive its environment, make decisions, and act towards achieving predefined goals. AI agents often use sensors and effectors and can exist in both virtual environments (like software agents) and physical forms.

Robot: A physical entity that uses AI to interact with the physical world. Robots typically combine AI systems and agents with hardware components such as sensors, actuators, and control systems to perform tasks ranging from simple, repetitive actions to complex, adaptive operations.

For this paper, the term **"intelligent systems"** was chosen as it effectively encompasses all of the above, highlighting their shared characteristic of utilizing AI to perform tasks.

Accountability, Responsibility, and Liability

As these terms are becoming more relevant, they need further definition. This increased relevance is due in large part to recent legislation emphasizing senior management's accountability in and ownership of cybersecurity, safety, and third-party risks (e.g., NIS2, DORA, SEC, CRA). This is highly relevant in the context of intelligent systems risk management and the extent to which leaders can be held responsible and/or accountable for risk identification, mitigation decisions, and the results of remediation actions. Senior leaders are increasingly becoming more responsible for AI assurance.

Accountability entails higher risk than responsibility, and both carry legal obligations (liabilities).

Accountability: Accountability implies oversight and means being answerable for the outcomes of tasks. It implies a higher level of ownership, where an individual or group is held liable for the success or failure of specific actions or decisions. Accountability is typically singular and involves reporting on results and justifying decisions at the highest level.

Responsibility: Responsibility refers to the duty to complete a task or adhere to a role as agreed. It involves being assigned specific obligations, and accepting them with the commitment and expectation that these will be fulfilled as agreed. Responsibility is often shared among team members and can involve daily tasks and broader objectives.

Liability: Liability refers to the legal obligation to compensate for any harm or damages caused by one's actions or decisions. It establishes a legal framework for holding individuals or an organization's management board accountable for their responsibilities and the consequences of their actions.
In essence, responsibility means "commitment to a job," while accountability means "being answerable for the outcome of the job's function." Liability ties these two concepts to legal consequences, ensuring that the actions of people holding accountable and responsible roles are bound by potential legal repercussions. This paper focuses on accountability, while responsible behavior and awareness of liability are expected.

Understanding the difference between responsibility and accountability can have material implications on a personal level. While risk management should not be guided by possible legal consequences, risk behavior is often influenced by short-term monetary benefits. The new legislation adds the important perspective of long-term risk management anchored in personal accountability for decisions. This requires a move away from short-term benefits towards more secure and sustainable solutions and implies a huge cultural shift.

Assess the Auditor

An in-depth literature review by Springer [12] highlights that auditing methods must improve alongside the technology being assessed by these methods. Auditing intelligent systems is a multi-disciplinary approach where different methods and individuals' skill sets must work together and complement each other. Effective auditing must merge technology-oriented, process-oriented, and business-oriented approaches.

Figure 2: Audit of Intelligent Systems (adapted from Springer [12])

The assessment of intelligent systems is an increasingly complex process that requires expertise from diverse fields such as audit methodology, computer science, business, philosophy, psychology, industry specialists, and legal studies, as well as thorough experience with compliance, risk, and governance.

The operational and organizational independence of auditors is essential, and cross-sector collaboration is recommended. An interconnected and structured assessment approach is recommended when auditing intelligent systems' business, environmental, and social impacts.

There is a significant need to evaluate the competence, skills, experience, qualifications, and the use-case-relevant business knowledge of AI auditors themselves to ensure their efficiency, effectiveness, and honesty as audit professionals. In addition, potential risks associated with personal biases, lack of knowledge, and conflicts of interest must be assessed to determine impartiality.

Regulatory bodies and professional organizations must play a crucial role in monitoring and protecting audit practices and setting standards of acceptable AI audit procedures to ensure valid results. Legally

binding non-disclosure agreements must exist to ensure that AI auditors can access all necessary evidence and artifacts while keeping all knowledge gained confidential on a need-to-have, need-to-see, and need-to-know basis.

Due to the scope of the complexity of auditing an intelligent system, a dedicated team of competent auditors shall be selected by an accountable body and assembled to ensure the success of the audit and the validity of results. An Intelligent System Audit Team shall be led by an authorized Chief Audit Officer (CAO) or Chief Audit Executive (CAE) who oversees all AI system audit activities, interprets results, and documents findings as to their opinion or judgment of the trustworthiness of the intelligent system. Auditors shall not only be equipped with multi-disciplinary expertise and critical, investigative, and abstract thinking abilities (and be able to conduct risk assessments beyond "compliance only") but also have the knowledge and experience to know and maintain adherence to current standards, regulations, legislation, and other best practices.

While the intelligent systems technology propels itself forward, it is important for regulators, developers, and deployers to acknowledge and agree that only mutual efforts will yield trustworthy and beneficial assessments of intelligent systems. Hence, it is important that the industry and regulatory bodies work closely together to upskill auditors to assess intelligent systems. Ideally, the development of a certifiable standard will be considered to assess the auditors and track their skills and performance.

AI Governance

Auditing intelligent systems requires a tailored approach within the context of the organization, which considers the criticality and all applicable organizational specifics of each system when planning, conducting, and documenting the audit. The audit team must apply critical, investigative, and abstract thinking accordingly when designing the audit program (multi-site, cross-divisional, and, in some cases, across legal entities and borders/time zones).

To define audit scope, goals, and success measures (and ensure transparency) it is essential to agree upon and use measurable intelligent systems metrics like key performance indicators (KPIs), key risk indicators (KRIs), or other applicable measurements.

This paper lists key assessment objectives for the trustworthiness of intelligent systems (in alignment with ISO/IEC 22989:2022 [13]). This is a rapidly evolving field both in technology and audit practices. As such, this list must be adjusted by the audit team for each individual audit when defining scope and audit methods. In addition, a risk-based approach combined with critical curiosity and investigative and abstract thinking must be applied at all times during the planning and fieldwork of such an assessment. The documentation of results must be agreed on with all stakeholders.

- **Transparency**: Visibility of all assets in scope and end-to-end traceability of all physical and digital components.
- **Explainability**: The intelligent system or its designers must be able to explain all decision-making processes to stakeholders (including end-users and auditors), if this feature is required/included. The National Institute of Standards and Technology (NIST) proposes the following four principles [14] for Explainable AI (XAI): explanation (reason for output), meaningful (understandable to the intended end-user), explanation accuracy (an explanation correctly reflects the reason for generating the output), and knowledge limits (system only operates under conditions for which it was designed and/or when it reaches sufficient confidence in its output).
- **Predictability**: The outcome or behavior of the intelligent system must be predictable. Depending on the risk associated, this might be a range of acceptable outcome or behavior or a very tightly regulated or even zero-tolerance scenario.
- **Controllability**: Ensuring controllability is a complex task to which utmost attention should be paid. It should be explored if the intelligent system could find ways to escape controllability. If an intelligent system is designed to function without human oversight, clearly regulated and documented guardrails must be installed.
- **Reliability**: The intelligent system that is assessed must prove to be reliable to deliver the quality and quantity of outcome or behavior for which it is designed.
- **Fairness**: Ensure the intelligent system has been trained on a diverse dataset that represents multiple aspects, including demographic groups, socioeconomic backgrounds, and varied viewpoints, to promote equitable outcomes.
- **Bias**: The intelligent system must be tested rigorously to ensure that it is fit for purpose in the intended context/use. Outputs of the intelligent system should not discriminate or be influenced by any unethical social conditions.
- **AI Resilience**: The system must be robust, resilient, and its plasticity (continuous learning) must be controlled and controllable. Depending on the criticality of the intelligent system, it must not only be fit for one purpose but also tested under changed parameters and/or in unfamiliar environments and provide reliable results evidencing AI Resilience.

Additionally, the following points are important aspects to consider:

- **Interpretability**: This addresses the factor that humans can understand the output.
- **Accountability**: Clear lines of ownership and responsibility for intelligent system behavior, outcomes, and decisions (regardless of their digital or physical impact). A shared responsibilities model can be established to visualize impacted areas.
- **Privacy-by-design**: The incorporation of privacy safeguards shall be evident from the initial design stage throughout the intelligent system's development lifecycle.
- **Security-by-design**: Secure development strategies must have been applied throughout the development lifecycle, and all code and libraries must have been verified.

The more advanced or critical an area of operation is, the greater the need for explainability and validation. Especially with LLMs and other pattern-producing systems, explainability is a key factor in understanding if the paths to a certain output are correct, are possibly wrong, or if they are simply beyond human comprehension. Assessing this type of AI governance risk (or inability to govern a risk that cannot be comprehended by a developer or their auditor) must also be documented within audit outcomes.

Applicable Laws, Regulations, and Standards

Knowing and ensuring compliance with applicable laws, industry-specific standards, and internal organizational policies is crucial. This is achieved through implementing an effective governance that addresses company-wide, innovative risk management and business continuity. An intelligent system must be auditable to evaluate compliance and risks beyond compliance [15].

The Cloud Security Alliance (CSA) offers a comprehensive resource hub on AI Governance & Compliance [16]. This hub was created by the CSA Workgroup AI Governance and Compliance and is regularly updated by volunteers. It provides a vast collection of documents covering diverse AI topics across various regions. It is free and fully searchable.

Use Cases as the Starting Point

With the help of the intelligent system owner, the audit team[1] must be introduced to the use case(s) to facilitate the understanding of the intelligent system's purpose and context. In a conventional audit that confirms compliance, a system is judged against established requirements, metrics, and controls. Those requirements are defined in laws, regulations, standards, organizations, and policies, and they might differ across regions and industries. It is the assessor’s responsibility to be familiar with all applicable compliance requirements, industry standards, and best practices. The goal of an assessment beyond compliance is to identify novel challenges that are not (yet) covered by regulatory bodies, as the current speed of innovation outpaces the regulatory capabilities. This paper is intended to provide considerations to enhance the trustworthiness of intelligent systems that aim at pushing the line of what is possible. Therefore, the use case guides the establishment of metrics adapted to the specific intelligent system from a perspective of an assessment beyond compliance.

[1] Audit Team: this could be an internal audit team, any kind of external audit team (including second and third parties), or an audit team from a regulatory body.

Also, the use case shall define the planned and/or (legally) required degree of human intervention to achieve the desired outcomes or behavior also in the face of an edge-case scenario. There must be clear guidelines regarding human oversight [15]. These elements are known as human-in-the-loop (active human oversight), human-out-of-the-loop (no option of human override), and human-on-the-loop (human is in a monitoring role and can take over control).

Auditors and end-users must acknowledge that, depending on the criticality of the system, severe incidents could occur if the use case is missing the right level of human involvement. There might be use cases where it is safer to specifically exclude human oversight. An example for such a use case is certain hydro infrastructure where decisions need to be made much faster than any kind of human oversight would allow. Also, it must be assessed if an "opt-out" or "opt-in" option for human review is implemented, and if a human can challenge the decision of an intelligent system. These considerations are an integral part of the initial use-case considerations and will impact audit results if not planned, documented, and implemented adequately.

A well-defined scope for the audit is also crucial. Audit planning must include the roles, responsibilities, and required availability of the auditors and key stakeholders. Knowledge of relevant regulations, binding norm requirements and company policies is required. Auditor and end-user safety, product quality and security, data confidentiality, integrity, and availability (CIA), incident management, compliance criteria, and business continuity must also be considered in the planning and reviewed during the audit.

Integrating this type of assessment within the corporate risk management processes and procedures ensures responsible intelligent systems use and encourages senior management's awareness of the intelligent systems deployed, for which they are accountable.

When assessing the use case and fitness-for-purpose of an intelligent system, assessors must have the ability to apply critical, investigative, and abstract thinking. By employing these methods, auditors can identify novel risks and possible non-compliance. Through this process of internal or external impartial review, recommendations for improvement can be made to enhance the security and trustworthiness of intelligent systems. This way, auditors can identify risks requiring mitigation by reviewing actual performance based on use cases, by challenging assumptions, and by providing a competent perspective of current intelligent systems operations. If audit recommendations are heeded, this can make a positive difference in the long-term security, safety, and trustworthiness of intelligent AI systems.

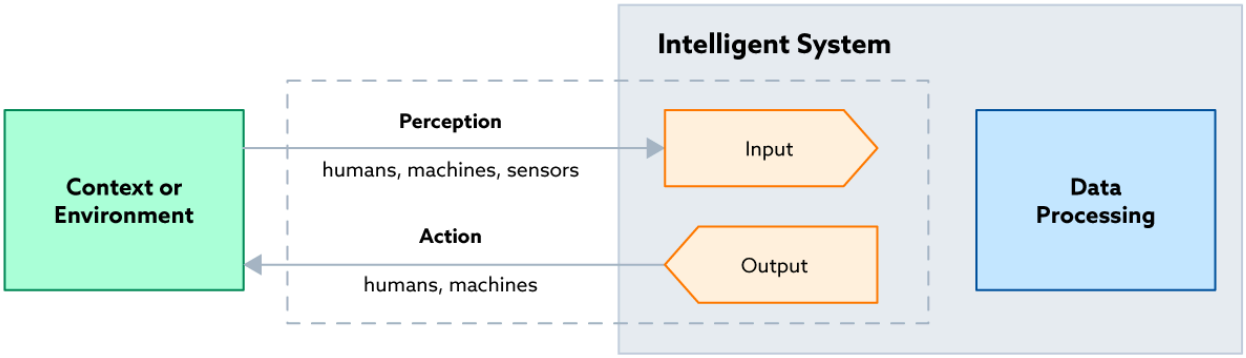


Figure 3: Intelligent System Definition (adapted from OECD)

The following sections delve into specific considerations that may or may not apply to a particular intelligent system but give an overview of concepts to consider.

Infrastructure for Intelligent Systems

Intelligent systems often handle sensitive data, making infrastructure assessments crucial and the implementation of advanced privacy methods (discussed below) necessary. Such an assessment is essential to gain insight in its support and influence on the intelligent systems' performance, including computational power, data processing speeds, and bandwidth, as well as its security, privacy, scalability, and sustainability. The infrastructure encompasses hardware components, networks, storage, and operating systems and software applications critical for developing, training, deploying, operating, and maintaining intelligent systems.

Implementing an effective cybersecurity management system [17] is paramount in ensuring the resilience of any intelligent system especially for those handling highly sensitive data. Existing frameworks from the NIST, as well as the Artificial Intelligence Risk Management Framework (AI RMF) [18] and GenAI RMF Profile [19], provide a foundation for managing intelligent systems' cyber and other key risks. Looking ahead, standards like the emerging Framework for AI Cybersecurity Practices (FAICP) [20] from the European Union Agency for Cybersecurity (ENISA) offer more specific guidance on securing intelligent systems throughout their lifecycle. An auditor must check for strict adherence (conformity) to these standards, frameworks, and best practices.

Energy efficiency and environmental sustainability have become vital concepts within the intelligent systems ecosystem, reflecting the need for greener technology solutions in the face of AI's significant energy demands. Laws regarding the carbon footprint are already in place in the European Union (EU). Auditors must also evaluate measures to optimize energy use, such as employing energy-efficient

hardware, utilizing renewable energy sources, and minimizing the environmental impact through green computing practices and responsible disposal of electronic waste. Staying updated regarding best practices, frontier solutions, and current regulations is the assessor's obligation. These combined efforts contribute to sustainable AI development and align with global climate change mitigation efforts and Goal 13 (Climate Action) of the Sustainable Development Goals (SDGs) published by the United Nations [21]. In order to ensure the long-term trustworthiness of intelligent systems, auditors must be able to think critically, investigatively, and beyond current regulations to make proactive, fact-based recommendations based upon their understanding of each system's use case, the IT architecture infrastructure, and their extensive knowledge of assessment criteria, including sustainability requirements.

Data Processing Unit

Data processing units are crucial for intelligent systems, encompassing CPUs, GPUs, Tensor Processing Units (TPUs), and edge processors found within Internet of Things (IoT) devices. CPUs handle general-purpose computing tasks, GPUs excel in parallel processing for complex calculations, and TPUs optimize machine-learning workloads. Edge processors enhance efficiency by processing data locally, reducing latency in IoT environments. Looking forward, emerging technologies like quantum computers promise exponential processing power by leveraging quantum mechanics for the improvement of certain computations. Biological computers explore utilizing biological materials for processing, offering the potential for bio-inspired computing. These innovations suggest a future where diverse computational paradigms contribute to advancing the capabilities of intelligent systems across various domains.

There are incredible advancements currently going on in this field and it's important to continuously update the personal knowledge to the latest insights, frontiers, and best practices. These new technologies offer plenty of room for critical thinking and proactive measures to ensure long-term trustworthiness of intelligent systems.

The Role of Sensors

All types of sensors, including visual, sound, touch, temperature, proximity, chemical compound and moisture detectors, play a crucial role in gathering data for intelligent systems. IoT is a highly dynamic field that quickly leverages advancements in sensor technology. This data is utilized by intelligent systems for various applications, as illustrated in Table 1.

Application	Description
Smart Cities	Sensor networks can monitor traffic flow, energy consumption, and environmental conditions, enabling intelligent urban planning and resource management.
Precision Agriculture	Sensors can provide real-time data on soil moisture, nutrient levels, and crop health, optimizing irrigation, fertilizer use, and overall agricultural productivity.
Advanced Robotics	Next-gen sensors can equip robots with a more nuanced understanding of their environment, allowing for safer interaction, improved agility, and autonomous operation in complex settings.
Personalized Healthcare	Biosensors can enable continuous health monitoring, early disease detection, and tailored treatment.

Table 1: Application of Sensor Data by Intelligent Systems

Advancements in sensor technology are poised for a significant leap forward through three key areas:

- **Miniaturization** (even using nano-technology)
- **Context-awareness**
- **Sensor fusion**

Miniaturization will enable the development of incredibly small sensors, vastly expanding their potential applications. Context-awareness will empower sensors to interpret their surroundings, yielding richer data and more insightful analysis. Lastly, sensor fusion will facilitate the integration of data from multiple sources, creating a more comprehensive environmental picture.

Benefits	Description
Improved Accuracy	Filtering out irrelevant data based on context leads to more accurate and actionable insights.
Enhanced Automation	Sensors can trigger automated responses based on context. For example, smart lights can turn on only when someone enters a room at night.
Reduced Power	By focusing on relevant data, context-aware sensors can operate

Consumption	more efficiently and extend battery life.
Personalized Experiences	Sensors can tailor their behavior to user preferences or environmental conditions, creating a more personalized experience.

Table 2: Benefits of Advancements in Sensor Technology

Sensor advancements introduce possibly underestimated or novel risks. Audits must assess combinations of sensor capabilities and the resulting potential harm from neglect, misuse, or intentional/malicious abuse.

Sensor owners and implementers must help assessors understand desired insights generated by sensor data in order to comprehensively understand the overall impacts and risks of these sensors. Also, the impact of a possible loss or faulty calibration of sensors to the intelligent systems under evaluation must be considered [22]. The accuracy and comprehension of this information is crucial for evaluating the trustworthiness of intelligent systems.

Table 3 gives an overview of some potential challenges; the list is not exhaustive and is subject to change over time.

Challenge	Description
Data Security and Privacy	Collecting and processing contextual data raises concerns about privacy and security. Robust data protection measures are crucial but realistically difficult to achieve.
Complexity	Developing and deploying context-aware sensor systems requires expertise in sensor technology, data analytics, and machine learning (ML).
Interoperability	Standardization is needed to ensure different sensors and systems can communicate and share contextual data seamlessly.

Table 3: Challenges Brought by Advancements in Sensor Technology

The next chapter sheds light on the complex management of data in intelligent systems.

The Role of Data

In today's world, where data [23] is the new currency and key to business success, especially in intelligent systems, understanding global data regulations, data quality, and the challenges and opportunities surrounding data (Figure 4) is crucial. These factors can "make or break" a business.

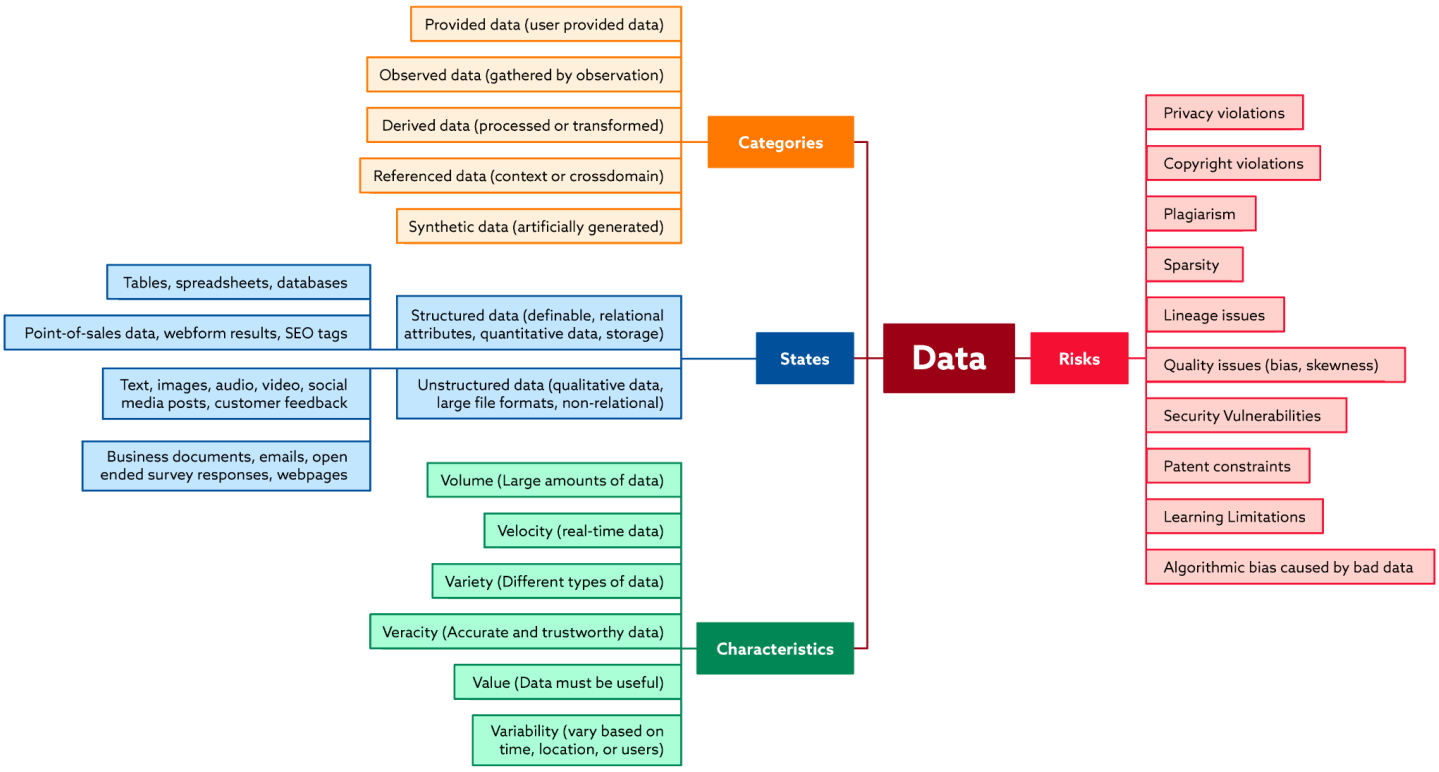


Figure 4: Role of Data in Intelligent System

The Role of Supply Chains

When reviewing an AI component provided by a third party[2], consider risks such as data privacy, confidentiality, integrity and availability, bias and discrimination, compliance, performance, integration, lack of explainability, AI-enabled cyber-attacks, and regulatory and compliance risks. To manage risk, intelligent system components should be traceable with a Software and AI Bill of Materials (SAIBOM). Recent events demonstrate the need for effective liability risk management of providers of software or hardware components and/or cybersecurity services. These risks shall be included in the supplier risk

[2] A third party is a supplier other than the developer or one of their suppliers.

management. The EU and other regulators require third-party risk management and, as such, actual vendors must be risk-assessed and their criticality determined. Intelligent systems assessors must extend their focus to the IT and third-party risk management of the company, possibly including assessments of partner vendors and/or critical suppliers with a focus on cross-border privacy regulations (if applicable). Third parties should be contractually mandated to disclose vulnerabilities and incidents within an agreed timeframe, with substantial monetary penalties for delays or failure to disclose.

Ensuring transparency and traceability is crucial to maintaining trustworthiness and the levels of these must be assessed during the assessment.

Data and Privacy

Various global regulations govern data, often with extraterritorial effects. For instance, the General Data Protection Regulation (GDPR) [24] applies to most personal data from EU entities, even if processed outside the EU. A Data Protection Impact Assessment (DPIA) [25] helps companies to identify non-compliance and fix gaps. Similarly, the California Consumer Privacy Act (CCPA)[26], the California Privacy Rights Act (CPRA) [26], and the Swiss Federal Act on Data Protection (FADP) [27] impose comparable data use constraints.

In a globalized world, the different legislations have to be navigated carefully; explicit consent is often required and any exceptions to applicable privacy laws must be explicitly accepted by the data subject. The "right to be forgotten" (GDPR) and "mandatory to remember" (data retention laws in regulated industries and the EU AI Act) are contradictory concepts, complicating data privacy in intelligent systems. Assessments must ensure the company has a transparent, legally sound consent management system that documents explicit and informed consent, especially where data retention is mandatory or non-avoidable (e.g., used as training data) and non-compliant with GDPR.

Data and Copyright

Data used for training intelligent systems can raise copyright concerns because the data might include external intellectual property and copyrighted material, and the legal line on fair use for training is still evolving and therefore not explicitly discussed here. Please refer to the CSA AI Resource Hub [16] for current publications. This creates a complex challenge, as AI-generated content from one system can become training data for another. Once copyrighted material enters the training loop, it becomes difficult to disentangle it from the “shared knowledge” of intelligent systems. Hence, an assessment must check the source of data and compliance with applicable laws to prevent any copyright infringements and to have a process in place that supports removal of copyrighted content (e.g., the notice-and-takedown system of the Digital Millennium Copyright Act (DMCA) [28]).

Data Sources

An upcoming issue is the potential bottleneck of available high-quality[3] training data that might become scarce as early as 2026 [29], [30]. However, these projections do not consider the increase in available data from sources like mass surveillance, automated vehicles, and the world of IoTs, as well as the production of synthetic data.

The black market [31] for data thrives on illegally obtained datasets, which include personally identifiable information (PII), financial data, medical records, and login credentials, posing significant risks to both individuals and businesses.

Transparency into "data pedigree" will become a relevant topic in the assessment of intelligent systems. Industries must prepare to demonstrate in-depth and trackable data management. This helps ensure data lineage as the data moves through the system, from its source to its end location, and any changes made to it along the way.

Organic Versus Synthetic Data

Organic data offers authenticity by reflecting real-world scenarios and behaviors, capturing nuanced and varied patterns trusted for their accuracy and relevance. However, it faces challenges, including limited availability in specialized domains, privacy risks with sensitive information, and high costs in obtaining, storing, and managing. In contrast, synthetic data excels in scalability, enabling precise manipulation while preserving privacy. Yet, it may lack the realism to fully capture organic complexities and may train an intelligent system with pre-existing, possibly even unintended, patterns.

A combination of using synthetic data for pre-training and organic data for fine-tuning has proven to even outperform under certain circumstances [32]. In an audit, the data provenance and the reasoning behind the use of organic and synthetic data must be transparent, explainable, and purposeful.
Auditing companies that generate and sell synthetic data requires a deep understanding of their generation methods. Unintentional or intentional patterns in this data can lead to unforeseen and potentially harmful behavior in intelligent systems.

As this field is evolving and encompassing new advancements, it's important for any assessor to be up to date on the latest advancements, best practices, standards, laws, and regulations.

[3] High-quality data consists of books, scientific papers, Wikipedia, filtered web pages, and so on.

Data Quality

The quality of data refers to its accuracy (free from errors), completeness (containing all expected attributes), consistency (uniform format and values), reliability (consistently trustworthy), timeliness (up to date), validity (correctly representing the real-world object or event), and relevance for the intended use. High-quality data is crucial for making informed decisions, ensuring reliable outcomes in analyses, and fostering trust in data-driven systems. Factors influencing data quality include collection methods, storage practices, and validation processes, which aim to minimize errors, biases, and inconsistencies. Continuous monitoring and maintenance are essential to uphold data quality over time, ensuring it remains fit for purpose and supports meaningful insights. Maintaining data quality is crucial for organizations relying on data for decisions and analysis.

Data Storage

Data storage solutions are critical for intelligent systems, encompassing various technologies to manage and process vast amounts of information efficiently. Internal storage options like Solid State Drives (SSDs)[4] and Hard Disk Drives (HDDs)[5] provide high-speed access to data stored locally, while Random Access Memory (RAM)[6] enables rapid data retrieval for immediate processing needs. Cloud storage solutions offer scalable and accessible repositories for storing and analyzing large datasets, enhancing flexibility, and facilitating the collaboration of intelligent systems across distributed environments. Within an audit for intelligent systems, it's important to check for sufficient resources to maintain performance, security, and safe data storage capacity.

As data volumes surge, researchers are exploring advanced storage methods. Holographic data storage [33] shows promise for high-density storage with rapid retrieval, while DNA-based storage [34] offers long-term digital storage solutions. Such novel data storage must be critically assessed in an audit to make sure it is suitable for the business case.

[4] SSDs are faster than HDDs and consume less energy.
[5] HDDs are older and slower technology compared to modern alternatives.
[6] RAM is used for data that is actively processed.

Networking, Connectivity, and Communication Interfaces

In cybersecurity, emerging AI-driven security solutions and advancements in encryption methods, including post-quantum cryptography [35], will play crucial roles in overcoming future challenges.

Connectivity and communication interfaces are crucial for intelligent systems, encompassing a wide range of technologies to ensure seamless data exchange and robust cybersecurity. These include traditional networking methods such as WLAN, Ethernet, and Cellular (4G/5G). Short-range wireless communication protocols such as Bluetooth, Bluetooth Low Energy (BLE), and RFID are also integral. Additionally, IoT-specific protocols like LoRaWAN, Zigbee, Z-Wave, NB-IoT, and Sigfox provide various options for low-power, long-range, and short-range communications.

To meet specific needs, like machine-to-machine communication and high-performance data exchange, industrial and real-time communication protocols such as OPC UA, DDS, and CoAP are utilized. Each of these systems introduces its own cybersecurity challenges, from vulnerabilities in wireless communication to securing data transmitted over constrained networks. Managing these challenges is essential to ensure the integrity, confidentiality, and availability of data in AI-driven applications.

Looking forward, the development of 6G networks promises improved connectivity and even lower latencies. Advancements in quantum communication and novel IoT protocols are expected to further transform the landscape.

Fog and Cloud Computing

Fog computing nodes, strategically placed closer to data sources like sensors and actuators, serve as intermediaries between the network edge and cloud data centers. These nodes, with their processing power and storage capabilities, perform local data analysis, filtering, and pre-processing, thereby reducing latency and bandwidth usage. Acting as intelligent gateways, they determine which data requires cloud-level processing and which data can be handled locally, optimizing overall cloud load and ensuring real-time application performance. This integration of fog and cloud computing enhances efficiency and enables continued operation even during internet disruptions.

Looking to the future, fog computing is expected to play a pivotal role in emerging technologies such as autonomous vehicles, smart cities, and industrial IoT. The integration of intelligent systems for predictive

analytics and edge applications[7] at the fog level will further enhance decision-making capabilities. Additionally, advances in cybersecurity measures, including the standardization of the Ascon [36] family for lightweight cryptography by the NIST, are anticipated to address the unique challenges of fog computing [37].

The integration of AI into fog and cloud computing systems introduces a new layer of complexity in maintaining data integrity and quality, making auditing processes more critical than ever. Auditing AI systems within this context involves not only verifying the quality of the data used but also evaluating the algorithms and models employed at both the fog and cloud levels. This requires continuous monitoring and assessment to identify biases, ensure compliance with regulatory standards, and validate the overall reliability of AI outputs. As fog computing nodes take on more responsibilities in data pre-processing and local analysis, auditing must also extend to these distributed environments, ensuring that the decisions made locally align with organizational objectives and ethical considerations.

Software Components

Software components are essential for the operation and management of intelligent systems. Virtualization software facilitates efficient resource allocation and management across virtual machines. Operating systems provide foundational support for hardware and software interactions. Middleware facilitates communication between disparate applications and systems. Application software and development environments enable the creation and execution of algorithms of intelligent systems and applications. Data management tools ensure efficient storage, retrieval, and processing of large datasets. Monitoring and logging tools track system performance and security. Container orchestration platforms streamline the deployment and management of intelligent system applications across distributed environments. Together, these components form the backbone of intelligent system infrastructure, supporting scalability, efficiency, and reliability in complex computing environments.

[7] Software that operates on devices at the network edge, processing data locally to enable faster response times and reduce reliance on centralized cloud resources.

Algorithms, Training Methods, and Models

Intelligent systems rely on three key components: algorithms, training methods, and models. These components rely partly on each other and exhibit intertwining dependencies without clear boundaries.

Algorithms define the core learning method as a blueprint for how a model learns from data. Training methods encompass the entire training process, including choosing the algorithm, preparing data, and selecting the most suitable method based on project goals. The training method will influence the choice of the model, which embodies the acquired knowledge and is used for predictions on new data through inference engines. Commercial or open-source frameworks typically offer a range of functionalities, libraries, and APIs that facilitate intelligent systems' development, training, evaluation, and deployment.

Maintaining the reliability and efficiency of these key components is crucial in production environments to ensure effective, efficient, and trustworthy intelligent systems.

More literature on these topics and their interdependencies can be found in textbooks, online courses, scientific publications, and discussion forums. A detailed understanding of each topic is required to fully understand the dependencies and have the ability to assess an intelligent system from a critical, investigative yet unbiased perspective.

In the past, there were several issues with data, algorithms, training, and models exemplifying the potential pitfalls of intelligent systems in decision-making [10]. Regulations like the EU AI Act (Art. 5) [38], [39] and Colorado Senate Bill 24-205 [40] are taking steps to address these concerns by labeling high-risk applications and encouraging responsible use. This emphasizes the importance of diligent adherence to compliance and the deployment of intelligent systems that are trustworthy even beyond current compliance standards by assessing systems critically and suggesting proactive measures to reduce risk.

Algorithms and Models

Choosing the right algorithm and model involves various key considerations based on the specific problem and available data. This quickly becomes quite technical, but it’s important for an auditor to have a good grasp of the underlying technology to be able to assess the risks of an intelligent system.

Also, the assessment of the computational resources to balance both the algorithm’s and model’s complexity with available computing power is required. Factor in the need for scalability, robustness to noise, and how well the algorithm or model handles missing data or poor data quality.

Human in the loop (HITL) is a collaborative approach in machine learning and intelligent systems where humans actively participate in the training and evaluation of ML models, providing valuable guidance, feedback, and annotations. Through this collaboration, HITL aims to enhance the accuracy, reliability, and adaptability of ML systems, harnessing the unique capabilities of both humans and machines. It is also a crucial concept for operating ML models and/or intelligent systems and can be legally required depending on the criticality of the intelligent system.

The following approaches cater to diverse applications and challenges, expanding the toolkit for developing intelligent systems:

- **Supervised Learning Algorithms:** These algorithms learn from labeled data, where the desired output is provided. Examples include decision trees, linear regression, and support vector machines.
- **Unsupervised Learning Algorithms:** These algorithms discover patterns in unlabeled data, where the data has no predefined outcome. Examples include clustering algorithms like K-means and dimensionality reduction techniques like Principal Component Analysis (PCA)[8] or autoencoders[9].
- **Semi-Supervised Learning:** Combines labeled and unlabeled data to improve learning accuracy and efficiency.
- **Transfer Learning:** Adapts knowledge from a source task or domain to improve learning and performance in a target task or domain.
- **Online Learning (Incremental Learning):** Updates models continuously as new data becomes available, adapting to evolving patterns and trends.
- **Multi-Task Learning:** Simultaneously learns multiple tasks to improve generalization and efficiency across related tasks.
- **Active Learning:** Guides the selection of new data samples for labeling to improve model performance with minimal labeled data.
- **Imitation Learning (Learning from Demonstrations):** Learns from expert demonstrations or behavior examples to replicate desired actions or behaviors.
- **Reinforcement Learning Algorithms:** These algorithms learn through trial and error interactions with an environment, receiving rewards for desired actions. Examples include Q-learning and Deep Q-Networks.
- **Natural Language Processing (NLP) Algorithms:** A subfield of machine learning focused on processing and analyzing human language. Examples include sentiment analysis, machine translation, and text summarization.
- **Computer Vision Algorithms:** This subfield focuses on tasks related to image and video analysis. Examples include object detection, image classification, and facial recognition.
- **Evolutionary Algorithms:** Problem-solving techniques inspired by natural selection, where candidate solutions evolve through processes like reproduction, mutation, and selection to find an optimal solution.

[8]PCA is a technique used to reduce the number of variables in a dataset while retaining its essential information.
[9] Autoencoders are neural networks designed to compress data into a lower-dimensional representation and then reconstruct the original data from this compressed form. They are versatile tools used for tasks such as dimensionality reduction, anomaly detection, and data generation, leveraging their ability to capture complex patterns and relationships in datasets.

These cutting-edge approaches are pushing the boundaries of what's possible and opening doors for new applications:

- **Federated Learning Algorithms:** Train models collaboratively across multiple devices or servers, without sharing the raw data itself. This approach is beneficial for privacy-preserving applications and distributed data scenarios.
- **Generative Adversarial Networks (GANs):** Involve two competing neural networks, a generator and a discriminator. The generator creates new data (e.g., images, text) that the discriminator tries to distinguish from real data. This competition leads to highly realistic generated content.
- **Self-Supervised Learning:** Uses pretext tasks to generate labels from input data without human annotation, enhancing model learning.
- **Meta-Learning (Learning to Learn):** Improves model learning and adaptation by learning how to learn from different tasks or domains.
- **Few-Shot Learning:** Aims to train models on very limited datasets, enabling them to learn from just a few examples. This is particularly valuable for tasks where acquiring large amounts of labeled data is expensive or difficult.
- **One-Shot Learning:** Learns from a single example or a few examples to generalize to new, unseen instances.
- **Zero-Shot Learning:** Transfers knowledge across domains to classify objects without labeled examples from the target domain.
- **Explainable AI (XAI):** As machine learning models become more complex, the need to understand their decision-making process grows. XAI techniques aim to make these models more transparent and interpretable, allowing humans to better understand their reasoning.

This list is not intended to be exhaustive and may change, encompassing both mainstream and frontier approaches. Ensure you approach each audit with an open mind while remaining critical.

Frontiers

As the future of intelligent systems is explored, quantum machine learning and neuromorphic computing emerge as exciting frontiers with the potential to revolutionize the field. Quantum machine learning [41] offers the ability to tackle problems beyond classical capabilities, while neuromorphic computing [42], [43] seeks to create hardware inspired by the brain's efficient information processing. While both technologies are in their early stages, they hold significant promise for overcoming current limitations in processing power, potentially leading to breakthroughs in research and hopefully reducing the carbon footprint of intelligent systems. Hand in hand with new possibilities come new risks, and auditors might be confronted with uncharted territory.

Fine-Tuning and Validation of the Model

Fine-tuning and validation are crucial steps for ensuring an AI system's model effectiveness. Training exposes the model to existing data, allowing it to learn patterns and generate predictions or insights. Fine-tuning refines the model by adjusting pre-trained layers for specific tasks, enhancing its capabilities. Validation with separate datasets assesses the model's ability to generalize and prevents overfitting to training data, ensuring accurate outputs for real-world applications.

Overfitting and Generalization

Ensuring an AI-system model generalizes well to new, unseen data involves techniques such as splitting the dataset into training and testing sets, applying cross-validation, analyzing learning curves, using regularization methods, and adjusting hyperparameters. These methods help evaluate whether the model is overfitting or generalizing well and guide adjustments to improve overall performance.

Fine-Tuning

Fine-tuning involves adjusting a pre-trained model's parameters using a smaller, task-specific dataset to adapt it to a specific task or domain. This process requires fewer computational resources than pre-training and can be done on lower-grade hardware. Innovations such as Low-Rank Adaptation (LoRA) have made fine-tuning more efficient and cost-effective. Fine-tuning can also incorporate human feedback to better align the model with desired behaviors.

Validation

Validation evaluates how a trained model performs in predicting outputs for new, unseen inputs using a representative validation dataset. This process calculates metrics such as accuracy, precision, recall, or F1[10] score to provide an unbiased assessment of the model's effectiveness. Performance testing evaluates the model's ability to handle various workloads and real-time scenarios, identifying potential limitations and bottlenecks. It measures computational efficiency and response time, optimizing the model's architecture and implementation.

[10] The F1 score balances precision and recall, making it useful for evaluating model accuracy on imbalanced datasets by considering both false positives and false negatives.

Furthermore, heightened focus on ethics in intelligent systems has led to the integration of fairness and bias detection frameworks, such as IBM's AI Fairness 360 and the What-If Tool by Google, into the validation process, ensuring models are not only accurate but also equitable. The rise of XAI techniques is also enhancing model transparency and trust.

Model Robustness and Stability

Evaluating an AI system's model robustness involves assessing its ability to resist adversarial attacks, maintain performance with noisy or corrupt data, and generalize to diverse scenarios. Stability ensures the model can handle unexpected inputs, provide transparent decision-making processes, adapt to changing data distributions, and minimize risks and biases.

Ongoing Maintenance

Establish procedures for ongoing testing, updating, and refinement of the AI system's models, algorithms, and performance metrics for the AI system to ensure continued relevance, effectiveness, and trustworthiness.

Frontiers in Fine-Tuning and Model Validation

As the field of AI systems continues to evolve, emerging trends in fine-tuning and validation are redefining how models are optimized and assessed. One major development is federated learning, which enables the training of models across decentralized devices while preserving data privacy and addressing regulatory compliance, making it particularly valuable in healthcare and finance. Additionally, zero-shot and few-shot learning techniques are gaining traction, allowing models to perform tasks with minimal fine-tuning by leveraging extensive pre-training on diverse datasets. Such innovations can shorten the model adaptation cycle.

Moreover, the adoption of continuous learning systems is becoming more commonplace, allowing models to adapt to new data in real-time and helping them stay relevant in dynamic environments. The use of synthetic data or organic data, respectively, for fine-tuning can increase the efficiency of this process.

Actuators

Robots rely mostly on actuators to translate commands into physical actions. These include common rotary options like AC, DC, and stepper motors, as well as servos for precise movements. Linear actuators, such as pneumatic cylinders, are used for pushing or pulling actions. Various grippers, including mechanical, magnetic, and vacuum designs, enable robots to grasp objects, while wheels and tracks provide mobility. The choice of actuator depends on the robot's specific needs for force, speed, precision, and movement range.

The future of actuators is bright, with possibilities like artificial muscles mimicking biological movement and microfluidic technologies offering highly dexterous manipulation. These advancements hold immense potential for creating more versatile and efficient robots. However, potential risks include unintended consequences of malfunctions or misuse, and the ethical considerations surrounding increasingly human-like robots must be carefully addressed.

Power Supply

Powering intelligent systems varies depending on their physical embodiment. Robots and other embodied intelligent systems rely on batteries, which are portable but limited in capacity; AC power, which offers continuous power but restricts movement; or hybrid systems, which combine both for enhanced flexibility. Cloud-based intelligent systems or agents leverage the data center's power infrastructure, which may include renewable energy sources like solar or wind. Smart power management systems are crucial for optimizing consumption across intelligent systems.

For intelligent systems with physical embodiment (robots):

- **Batteries:** Portable and rechargeable but have limited capacity and require charging.
- **AC Power:** Direct connection to a wall outlet for continuous operation, but limits mobility.
- **Hybrid Systems:** Combine batteries and AC power, providing both portability and extended operation.

For intelligent systems without physical embodiment (cloud-based systems):

- **Datacenter Power:** Large-scale intelligent systems running in data centers rely on the overall power infrastructure of a data center, which may include a combination of grid power, backup generators, and renewable energy sources.
- **IoT and Other Small Devices:** These devices often rely on batteries, energy harvesting techniques, or low-power wireless charging. Power efficiency is critical due to their small size and often remote or

hard-to-reach locations. Strategies such as low-power chip designs, efficient sleep modes, and intermittent connectivity are essential to extend battery life and ensure reliable operation.

Additional considerations:

- **Power Management Systems:** Optimize power consumption for both embodied and non-embodied intelligent systems, ensuring efficient operation.
- **Wireless Charging:** Emerging technology for robots and other embodied intelligent systems, eliminating the need for cables and increasing mobility.

While this list covers the main power supply options for intelligent systems, it's important to remember that the specific needs will vary depending on the type of system, its processing power requirements, and its operational environment.

Power supply for critical infrastructure, including possibly intelligent systems, requires a multi-faceted security approach. Power grid vulnerabilities necessitate redundant power supplies and fault-tolerant grids to bolster resilience and prevent large-scale power outages. Additionally, shielding against Electromagnetic Pulses (EMPs)[11] and securing power management systems from cyber attacks is crucial.

User Interfaces

Ultimately, it is a leadership duty to foster a culture of accountability among developers, ensuring that innovations in user interfaces (UIs) are not only captivating but also adhere to the highest standards of trustworthiness. It is the assessor's obligation to ensure compliance and make proactive suggestions if needed.

User interfaces evolve to meet diverse needs, from traditional displays and panels to voice interfaces and mobile apps for hands-free use. Touchscreens and gesture recognition further streamline interaction.

The future UI landscape, including Augmented Reality (AR), Virtual Reality (VR), Brain-Computer Interfaces (BCIs), and haptic feedback[12], is geared towards immersive and accessible experiences. AR/VR offers real-time data overlays for surgeries, while BCIs enable thought-based control, which is crucial for users with limited mobility. Enhanced engagement from haptic feedback and the emphasis on accessibility considerations mark a revolution across various sectors.

[11] A powerful burst of electromagnetic energy that can damage electronic equipment. EMPs can be caused by natural phenomena like lightning strikes.
[12] Haptic feedback in user interfaces refers to the use of tactile sensations to communicate information to the user through physical touch.

AI-integrated UIs promise a shift to personalized experiences by adapting in real-time to user needs, which could notably improve accessibility by customizing presentation and interaction to fit individual requirements. This move towards personalization will serve to heighten convenience and inclusivity.

Despite the benefits, hyper-personalization bears risks, like reduced diversity in information exposure, which could affect creativity and collaboration. A balanced, nuanced approach is essential to maintain UI experience diversity alongside personalization's advantages.

Future UI advancements bring critical security, safety, and privacy considerations. The extensive data collection by BCIs demands robust protection and raises ethical questions. Moreover, AR/VR's immersive nature requires heightened safety measures to maintain situational awareness, and the interconnected UI landscape presents new vulnerabilities, highlighting the need for stringent defenses against potential threats.

From an auditor's perspective, assessing new UIs with developers is crucial for thoroughly understanding their potential benefits and inherent risks, including safety, privacy, and security concerns. An auditor must be able to methodically evaluate the risk profile of different UIs and deduct corresponding guardrails to maintain safety, security, and privacy. Auditors play a critical role in ensuring that the pursuit of appealing UIs do not overshadow the imperative to protect users, advocating for a design philosophy that balances user engagement with ethical responsibility.

Control Systems

Intelligent systems utilize various control systems to translate decisions into (physical) action. Centralized control, while simple, suffers from vulnerability.

- Decentralized control systems enhance robustness and adaptability by distributing control.
- Hierarchical control systems organize control systems across multiple levels, from high-level planning to low-level execution, enabling complex behaviors through layered decision-making.
- Behavior-based control systems excel in pre-defined situations but lack flexibility.
- Hybrid approaches leverage the strengths of combining different controls, offering more balanced solutions.

Future control systems promise even greater sophistication, incorporating adaptive learning and real-time optimization through techniques like evolutionary algorithms (self-optimizing algorithms), which optimize control systems through simulated evolution. This could lead to robots with superior decision-making and seamless and dynamic environmental interaction.

However, ensuring trustworthiness in these complex systems is paramount. Decentralized control at the edge of hybrid systems could mitigate risks from unintended autonomous actions and vulnerabilities from interconnectedness. Careful design and robust security protocols are crucial for responsible development.

Safety Systems

Critical thinking and rigorous risk assessments are essential for the development of trustworthy intelligent systems. Safe operation is particularly important for intelligent systems, as they rely on software safeguards to prevent unintended consequences. These can include fail-safe mechanisms that automatically shut down operations in critical situations or algorithms designed to detect and avoid malfunction in decision-making.

For embodied agents and robots, additional physical safety measures are crucial. Collision detection and avoidance systems, emergency stop mechanisms, and physical barriers like enclosures all work together to minimize the risk of accidents. Redundancy in critical components ensures continued safe function even if one part fails.

The following list summarizes the safety measures for intelligent systems:

- **Physical Safeguards:** These include mechanical barriers, enclosures, collision detection and avoidance, and limitations on speed or force to prevent physical harm.
- **Sensor Fusion:** Combining data from multiple sensors creates a more comprehensive and context-aware picture of the environment, aiding in safe navigation.
- **Fault Tolerance:** Designing systems with redundancy in critical components allows them to continue operating safely even if one component fails.
- **Safety Software:** Software algorithms can monitor system behavior and intervene to prevent unsafe actions. This could involve triggering emergency stops or limiting functionality. All sophisticated, intelligent systems should have a "kill switch," similar to manufacturing machines' emergency stop.
- **Human-Robot Collaboration Safety Protocols:** As robots work alongside humans, specific protocols and safety measures are needed to prevent accidents and ensure clear communication between humans and robots.

Ensuring ethical decision-making in autonomous systems and robots remains a challenge. Unforeseen consequences of (physical) actions or vulnerabilities arising from deeply interconnected systems could lead to safety hazards, data breaches, and privacy violations. OSHA [44] highlights many of the traditional risks and hazards that arise when working with robots, as well as some safety considerations.

Products with novel, innovative, and enhanced trustworthiness on all levels can make a competitive difference in tomorrow’s technical landscape. From the perspective of auditing intelligent systems beyond compliance, such systems could also help to make this world a more trustworthy place to be.

Security Systems

Security systems for intelligent systems are crucial to protect against threats and vulnerabilities. These systems include encryption to secure data transmission, access controls to limit unauthorized access, vulnerability scanning to identify any vulnerabilities in the system architecture, software code and system configuration, and intrusion detection systems to identify and respond to potential breaches. Offensive security systems and proactive security measures combined with regular security audits and timely updates ensure a high level of protection. Secure boot mechanisms and firmware integrity checks safeguard the system from malicious software. While the risk of introducing intelligent systems is widely discussed, human failures are still the most prevalent security issue. The cultural change needed to transform a crowd into a "human firewall" seems very difficult to achieve. Strategies such as continuous training, rewarding, secure behavior, and leadership endorsement can aid with this transformation.

Here is an overview of essential security measures for intelligent systems:

- **Encryption**: Ensures data confidentiality by encrypting it during both storage and transmission. Implementing dual encryption techniques prepares systems for resilience against future quantum computing threats.
- **Data Sanitization:** Secures the deletion of unnecessary or expired data, ensuring it cannot be recovered, even with advanced tools. This measure is particularly critical for protecting sensitive information.
- **Authentication:** Confirms the identity of users and devices requesting access, utilizing methods such as passwords, tokens, biometrics, or multi-factor authentication (MFA) to ensure that only authorized entities gain access.
- **Access Control:** Regulates access to system resources, assigning permissions based on predefined rules to prevent unauthorized access to sensitive data and functionalities.
- **Security Information and Event Management (SIEM):** Integrates and analyzes security data from multiple sources, providing a centralized platform for threat detection, incident response, and comprehensive security monitoring.
- **Intrusion Detection Systems (IDS):** Continuously monitor network traffic and system activities to identify through anomaly-based detection or signature-based detection, for example, and alert on suspicious behaviors that may indicate potential security breaches.
- **Continuous Monitoring:** Implements ongoing surveillance of system activities to detect anomalies and potential security incidents in real-time, enabling proactive defense measures.
- **Vulnerability Management:** Actively identifies and mitigates security vulnerabilities in system architecture, configuration, and software and hardware components through regular updates and patching.
- **Incident Response Planning:** Develops and maintains a structured approach to address and manage a security breach or cyberattack, aiming to handle the situation in a way that limits damage, reduces recovery time and costs, and maintains business continuity.

- **Disaster Recovery**: Ensures that there are plans and mechanisms in place for the recovery of critical data and systems in the event of a significant disruption, such as a cyberattack or natural disaster.
- **Physical Security:** Protects the physical hardware components of intelligent systems from theft, tampering, or damage, which is often an overlooked but vital aspect of security.
- **Security Awareness Training:** Educates users and operators of intelligent systems on potential security risks, best practices, and incident response procedures to maintain a trustworthy security posture.
- **Penetration Testing and Bug Bounty:** While PenTests are required by new EU regulations for certain critical industries, a bug bounty program is voluntary but can be of huge benefit to discover possible vulnerabilities before they are being exploited by malicious actors.

Last but not least, periodic audits are important to evaluate the effectiveness of security measures, ensure compliance with industry standards, and identify areas for improvement.

The rise of AI-driven (offensive) security tools and zero-trust architectures is shaping the future of intelligent system security. Technical security also promises advancements through self-healing systems that can automatically detect and patch vulnerabilities. Such advancements enable more secure systems that adapt to evolving threats in real-time. However, challenges remain; the interconnected nature of systems creates complex attack surfaces that are rapidly growing. Mitigating these risks also requires a multi-layered approach, combining advanced security technologies with ongoing vigilance, threat assessments, and, ideally, the improvement of cultural awareness of these risks.

Advanced Privacy Methods

This overview is essential for evaluating intelligent systems with stringent privacy requirements. For systems with minimal or no privacy concerns, this section may be less relevant and can be reviewed as needed.

Throughout the entire lifecycle of an intelligent system, especially those handling sensitive data, privacy-enhancing techniques are crucial. This section provides an overview of key methods at various development stages, summarizing both previously discussed concepts and those specific to privacy. Please note that some of these methods are still being refined. It’s an assessor's obligation to update the required knowledge beyond best practices and stay informed about frontier methods, their advantages and possible risks.

Infrastructure:

- **Data Residency:** Stores and processes data within specific regions or countries to comply with privacy regulations.

- **Decentralized Storage:** Distributes data across locations, minimizing the impact of a data breach or a natural disaster.
- **Hardware Security Modules (HSMs):** These tamper-resistant devices safeguard cryptographic keys for data encryption, adding an extra layer of physical security.

Data collection and preparation:

- **Data Minimization:** Collect only the data strictly necessary for the AI model's function.
- **Data Anonymization and Pseudonymization:** Transform data to remove or mask PII while preserving its statistical properties for training.
- **Differential Privacy [45]:** A mathematical technique that injects noise into training data, improving model accuracy while protecting individual privacy.
- **Federated Learning [46]:** Trains AI models on decentralized datasets without transferring the data itself. Each device trains a local model on its data and shares only the model weights, not the raw data.

Training and development:

- **Synthetic Data:** Artificial data that resembles real data but doesn't contain any actual personal information. This can be used to supplement real datasets for training and/or fine-tuning.
- **Homomorphic Encryption:** Allows analysis of encrypted data without decryption. However, computations are currently slower, making them less suitable for real-time applications or large datasets.
- **Privacy-Preserving Machine Learning Algorithms:** These algorithms are specifically designed to minimize the information revealed about the training data during the model creation process.

Deployment and operation:

- **Secure Multi-Party Computation (SMPC):** Allows multiple parties to jointly compute a function on their data without revealing their own data to each other. This can be used for privacy-preserving data analysis.
- **Confidential Computing [47] (Utilizing Trusted Execution Environments (TEEs)):** This technique isolates and protects data while it's being processed within the AI system's model. Even if the main system is compromised, attackers cannot access the data itself within the secure enclave of the TEE.

Auditors assess the appropriateness of the techniques used in intelligent systems by evaluating them against established standards, industry best practices, and regulatory requirements. If an auditor is not a subject matter expert (SME) , then the audit staff should be augmented.

Development and Debugging

Throughout the whole development lifecycle, a variety of tools enable the creation and refinement of intelligent systems. Special attention is to be paid to the supply chain as an insecure supply chain can introduce vulnerabilities to the intelligent system [48]. Collaborative development platforms offer code co-pilots, version control, and the ability to revert to previous versions, facilitating robust software engineering. Debugging tools help identify coding errors, while performance monitoring tools ensure optimal efficiency. Machine learning frameworks come equipped with tools and libraries tailored for developing intelligent systems, and simulation environments provide safe spaces to train and test AI system models.

Automated testing frameworks primarily focus on ensuring code functionality and catching bugs, with significant growth potential at the application level. Profiling tools allow developers to analyze code performance and pinpoint efficiency bottlenecks. Explainable AI (XAI) tools assist developers in understanding how models make decisions, which is crucial for effective debugging and performance optimization.

While future advancements promise even more sophisticated tools, there are potential risks associated with over-reliance on automation. This dependency could lead to unforeseen consequences, highlighting the importance of tool trustworthiness. The integrity of these tools heavily influences the trustworthiness of the resulting intelligent system, highlighting the need for thorough auditing and a deep knowledge of this field.

As intelligent systems evolve to become self-optimizing, the dependency on third-party suppliers may decrease. However, the more unpredictable a component is, the greater its inherent variability and, hence, the risk of unexpected behaviors or undetected malfunctions despite human oversight.

End-User Training and Documentation

Documentation for the End-User

Detailed end-user documentation for intelligent systems is essential for transparency, training, maintenance, compliance, and successful integration into business processes. The NIST AI RMF outlines the necessary elements for AI system documentation, including capabilities, limitations, use cases, technical data, performance metrics, security controls, user guides, maintenance, support, and best practices. Documentation should be clear, concise, audience-specific, adapted to different accessibility paths, and regularly updated. Key elements include:

- **User Guide:** A structured user guide assists end-users in effectively using intelligent systems. It should include a quick start guide emphasizing any risks (including a disclaimer, if necessary, and/or legal accountability), system architecture, data management for privacy and security, configuration guidelines, user interface descriptions, possible use cases, troubleshooting, a glossary, and support contact information with the possibility for end-user specific feedback. Clear instructions, detailed explanations, and visual aids help users quickly learn the system, and online training for a specific intelligent system could be offered or made mandatory before access to the system is granted.

- **Maintenance Guide:** The maintenance of intelligent systems requires, at minimum, continuous monitoring, security and incident response, version control (including outlining the possibility of reverting to a previous version if necessary), configuration management, regular maintenance (including proper data management for continuously learning systems), and compliance (including any updates to the SAIBOM). Maintain schedules for system upgrades, security checks, and data backups. Document technical troubleshooting and provide a technical support contact and an online form to disclose any technical issues with the intelligent system.

- **Release Notes or Change Logs:** Release notes or change logs summarize system changes (including bug fixes), updates (including new features), and improvements. They are relevant for regulated industries to make sure the intelligent system stays compliant and keep end-users informed about changed functionality and possible new training requirements.

- **API Reference Documentation:** API reference documentation helps developers use APIs associated with the intelligent system, facilitating development and integration. It details API design, structure, functionality, inputs, outputs, parameters, and possibly an API access code. APIs are crucial for standardized integrations and help with automation. Their access has to be controlled, especially if actions could be triggered.

- **Disclosures:** Disclose how intelligent systems use data and how generative AI generates content, ensuring transparency. Users must be informed if they interact with an intelligent system. This

disclosure is crucial for ethical and legal reasons to prevent misleading users and support the "human-in-the-loop" concept.

- **Transparency:** Transparency reports provide public access to technical aspects, like architecture and training data, and outline measures to address fairness, bias, privacy, and security. These reports demonstrate responsible and accountable intelligent system development, highlighting features and areas for improvement. This results in continuous improvement and allows end-users to suggest novel solutions.

- **Privacy and Confidentiality:** An intelligent system privacy and confidentiality notice informs users about data collection, use, protection, access, storage, retention, and their data rights. It helps users understand the particular application their data supports and fulfills legal obligations.

- **AI Usage Policy:** An intelligent system usage policy outlines guidelines for responsible and effective use, the intended data usage, system scope and purpose, and authorized users or the requirements to become an authorized user. It aligns use of the intelligent system with organizational values, goals, and legal requirements and includes security and safety protocols.

- **SAIBOM & Data Storage:** The bill of materials must include physical materials, software, and AI components. The physical storage location of data must be documented, as must the location of the audit trail and related logs.

- **Operational Status Monitoring:** Operational status monitoring of intelligent systems ensures service reliability and required performance by tracking usage data, response times, error rates, and availability. It helps identify and resolve issues quickly, ensuring performance standards are met, which minimizes downtime and maintains a secure state of operation.

- **Help Center:** A well-designed and capable help center with sufficient resources demonstrates a commitment to providing reliable and accountable intelligent system product support, building user trust, satisfaction, and loyalty. It provides comprehensive resources like user guides, tutorials, training, frequently asked questions (FAQs), troubleshooting tips, and technical support, ensuring users get the help they require promptly and professionally.

- **Responsible/Accountable AI Practices:** Responsible/accountable AI practices ensure ethical and sustainable development and use of intelligent systems, minimizing harm and maximizing benefits. Such practices include using diverse datasets, mitigating bias, ensuring transparency, protecting privacy, and providing human oversight.

- **Disclosure of the Copyrighted Works Used in Training Data:** Disclosing the training data source, including copyright limitations, ensures transparency and legal compliance, mitigating infringement risks. Demonstrating ethical and respectful use of third-party content demonstrates responsible and ethical business conduct.

- **Feedback Processes for End-Users and Impacted Communities:** Feedback processes are crucial for end-users and impacted communities to provide input on system outcomes and content quality to ensure fairness, accuracy, and overall benefit, leading to possible improvements. There are different feedback methods available, including surveys, user behavior analytics, user testing, and social media monitoring. Feedback obtained can be used to prioritize areas for improvement and enhance user experiences. Customer support reflects user experiences and difficulties, which is useful for identifying areas for improvement.

(Mandatory) Training and Training Record Keeping for the End-User

ISO 42001 [8] provides training guidelines for individuals involved in the development or use of intelligent systems, identifying necessary skills. Training should cover principles, methods, ethics, data protection, cybersecurity, and legal requirements. It must be structured, designed to be participative, and include evaluation and feedback. Documentation, including attendance, evaluations, and qualifications, must be maintained (sometimes even for compliance). Training resources should be accessible to diverse audiences, understandable, and regularly updated, and include online tutorials and documentation. Ideally, resources allow hands-on training (possibly in simulated environments or sandboxes), gamification, or other engaging ways to enhance the learning effect and ensure the required skills' reliability and applicability.

Webinars about the latest features are a great way to interact with the end-users and also allows them to get answers from experts. An accurate and up-to-date knowledge base with optional additional information can further be a helpful resource for end-users.

Community forums are an excellent way to not only get answers from other users but also to give developers an overview and spot common or severe issues. Collaboration platforms often provide useful information for developers or end-users who wish to gain a more in-depth insight into the application.

Ensuring all users of intelligent systems are adequately trained (with documented training histories) and regularly updated enhances reliability and mitigates unintentional misuse that could lead to harm and/or business losses. Adequate training in identifying knowledge gaps helps to improve work quality and also supports legal requirements and compliance. In train-the-trainer approaches, it is crucial to verify the technical and social skills of the trainer to ensure reliable transfer of knowledge.

Deployment and Monitoring

Auditing the deployment and monitoring phases of an intelligent system is critical for ensuring that the system performs as intended and can handle unexpected situations reliably and securely.

Continuous Evaluation: Implement real-time performance monitoring to track key objectives and detect drifts. Defined as deviations from the objectives (purpose) or goals of an intelligent system, drifts can lead to minor inaccuracies. However, in combination with self-healing abilities and evolutionary algorithms, drifts can also cause significant disruptions and pose a risk that is not even quantifiable at this point. Despite sounding mundane, continuous evaluation is one of the most critical concepts for maintaining the trustworthiness of intelligent systems.

Version Control and Documentation: Maintain thorough change tracking and detailed documentation to uphold system integrity and facilitate future audits. A reliable change management process with a clear audit trail enhances transparency and accountability.

Incident Response: Establish detailed procedures for addressing failures and security breaches. Whether the intelligent system is embodied or not, it must have an accessible "kill switch" or another emergency shutdown method, which can pose a challenge in itself. Technical and organizational measures must be in place to address incidents quickly and reliably. Additionally, compliance with cybersecurity regulations like NIS2 and SEC requires having prompt and proper disclosure and response protocols.

Cybersecurity: In the deployment and monitoring lifecycle phase, the most critical aspects are:

- **Access Control**: Implement strict access control mechanisms to prevent unauthorized use of the AI system. This ensures that only authorized personnel can interact with sensitive parts of the system, reducing the risk of malicious activities or physical actions. Extend access control to other digital entities and implement a zero-trust architecture, combined with a reset of access rights at predefined or even random time intervals.
- **Intrusion Detection**: Use proactive monitoring tools to identify and respond to malicious activity and anomalies. Early detection of intrusions is crucial for minimizing potential damage and maintaining system integrity and trustworthiness.
- **Vulnerability Scanning and Patching**: Regularly scan for vulnerabilities and apply timely patches. Identifying and promptly addressing weaknesses is essential for protecting the system from potential exploits and ensuring its continuous, secure operation. Vulnerability Disclosure Programs (VDPs), bug bounties, or even honeypots can reduce the risk of breaches, although such tactics may not be effective against Advanced Persistent Threats (APTs) and silent compromises.

Release Management in Regulated Environments: Adhere to strict standards for deploying updates or new versions of the AI system, ensuring safety and compliance with regulatory requirements. While compliance is legally crucial, it's essential to ensure that risk assessments go beyond mere compliance. Verify that the intelligent system performs as intended and can handle unexpected situations securely.

Decommissioning

Safe decommissioning of intelligent systems involves securely erasing all stored data to protect sensitive information, dismantling hardware in compliance with environmental regulations, and ensuring all access controls are disabled or deleted with proper notification, and any software licenses or third-party integrations are properly terminated. Proper process documentation is essential for records, accountability, and future reference. An auditor should assess any potential disassembly hazards, such as hazardous materials, unexpected behavior, or threats to business continuity.

This phase includes several critical components. First, auditors should verify that data migration processes are secured and that decision-making processes remain uninterrupted, with logic preserved or updated in successor systems to continue operational and business activities.

Emergency decommissioning requires robust business continuity plans. If business continuity is secured, the likelihood that an erroneous intelligent system can be (or actually is) switched off immediately is greatly enhanced.

Auditors must confirm that organizations have clear and actionable policies for securely retaining necessary data and disposing of obsolete data, ensuring that sensitive information is permanently deleted from all storage media (including biological storage options, which can present unique ethical challenges). Logs and metadata must be archived securely for future reference, legal compliance, and forensic investigations. Auditors should ensure that these logs are stored in a manner that allows easy retrieval while protecting the integrity and confidentiality of the data. Future challenges might also involve addressing ethical considerations in decommissioning intelligent systems.

Conclusion

The intention of compliance is trustworthiness. Due to the rapid evolution of intelligent systems, this paper suggests a risk-based, critical approach to better intelligent systems beyond the scope of regulatory compliance. Investigative thinking, curiosity, and the passion for continuous learning are prerequisites to assess intelligent systems holistically.

Intelligent systems are now integral to various sectors, from healthcare and finance to transportation and national security. This ubiquity underscores the need for auditors who are willing and able to assess intelligent systems beyond ticking checkboxes. Trustworthiness requires not just compliance but commitment and this is inherent to an assessor's attitude and ethical standards. Considering the unprecedented pace at which intelligent system technology advances, mitigating risks that regulators may not yet address can be considered as proactive compliance and facilitate future-proof intelligent systems.

One critical area of focus is the continuous improvement of auditing processes, which must evolve in tandem with technological advancements and emerging threats. However, achieving this ambitious goal is not something that any single person or entity can accomplish alone. It demands collaboration, bringing together cross-sector stakeholders to pool resources, share knowledge, and establish best practices that can be universally adopted.

The ultimate goal is to foster an inclusive and diverse environment where intelligent systems enhance human capabilities without compromising their safety.

The path to trustworthy intelligent systems lies in recognizing that trustworthiness is a shared responsibility. Collaboration to ensure that technology and auditing can reach beyond compliance helps increase trust in current and future intelligent systems.

Appendices

The following questions do not consider the full integration of different risk management systems within a company to provide an overview of all risks identified. It is highly recommended to maintain a central risk management and governance approach to assure that interdependent risks are managed, mitigated, or transferred, ensuring that a holistic response is possible in the case of an incident. The questions in the appendices follow the previously discussed chapters and are intended to spark curiosity, risk awareness, and investigative thinking.

Appendix 1: Audit Questions "Audit the Auditor"

AI System Capabilities and Processes:

- How does the intelligent system select and prioritize audit candidates?
- What data is used by the intelligent system to perform the audit?
- How is the intelligent system trained to identify specific types of risks and anomalies?
- How does the intelligent system process and analyze the data to detect potential issues?
- What human intervention is required throughout the audit process?
- How is the accuracy and efficacy of the intelligent system's conclusions tested and validated?
- What safeguards are in place to ensure that the intelligent system does not introduce bias into the audit process?

Audit Program Design:

- How is the audit program regularly reviewed and updated to adapt to changing risks and technologies?
- Does the audit program clearly understand the potential risks and opportunities associated with the use of intelligent systems within the organization?
- Is the audit program team equipped with the necessary knowledge and skills to assess the effective use and potential risks of intelligent systems within the organization?
- Does the audit program have a framework or methodology that is specifically designed to review the use of intelligent systems or a particular intelligent system?
- How does the audit program monitor the intelligent systems within the organization to ensure compliance with relevant regulations and standards?

Audit Team and Collaborations:

- How is the audit team composed, and what expertise do they have in different fields such as computer science, philosophy, psychology, legal studies, and industry specialties?
- What processes are in place to ensure that auditors are operationally and organizationally independent of the intelligent system being audited?
- How are cross-sector collaborations carried out to bring together diverse perspectives and expertise in the audit process?
- What methodologies, tools, and techniques are used to audit the environmental and social impacts of intelligent systems in practical scenarios?
- How are auditors evaluated, trained, and mentored to maintain their knowledge-based and ability-based skills?
- What protections are in place to prevent conflicts of interest and personal biases within the auditing team?
- What regulatory and professional organizations are involved in monitoring and protecting the audit practices of intelligent systems?
- How are non-disclosure agreements used to ensure auditors have access to necessary evidence while keeping the intellectual property confidential?
- How is the trustworthiness of the intelligent system being judged, and what criteria are used to determine the level of trustworthiness?
- What mechanisms are in place to ensure that the audit program is improving alongside the evolution of technology, and what steps are being taken to ensure that the audit methods are effective and efficient?

Appendix 2: Audit Questions "AI Governance"

Governance and Accountability:

- Is an auditing structure present in the organization?
- Is there a Chief Artificial Intelligence Officer (CAIO) role defined as responsible for the coordination and oversight of AI activities within the company?
- With which teams does the CAIO collaborate? Are there any external consultants involved and what is their role?
- Is a single point of contact for the external auditor established? Who is this?
- Is a clear structure of responsibilities established?
- Is AI governance in place and integrated into the risk management of the whole company, including business continuity aspects?
- Does senior management allocate sufficient resources for the development of trustworthy AI?
- Is there a clear ownership structure in place, and are the lines of shared responsibility and ownership defined for the behavior, outcomes, and decisions of the intelligent system?
- Has the company developed a framework for responsible use of intelligent systems?
- Has the company established policies regarding the acceptable and ethical use of intelligent systems?
- How are policies regarding acceptable and ethical use of AI documented and communicated to appropriate individuals throughout the company?
- How does the company monitor compliance with policies regarding acceptable and ethical use of intelligent systems?
- Does the company have a process to monitor and track the use of intelligent systems throughout the company?
- Does the company have the possibility to monitor and track access to the intelligent systems by third-party service providers?

Risk Assessment:

- Has the criticality of the AI system been assessed, and has the audit approach been tailored for assessing the system's criticality and organizational specifics?
- Was a risk assessment conducted that took into account the criticality of the system?
- Are risk assessments planned at predefined intervals according to the criticality of the system?
- Are there "accepted risks"? Are they signed off on by senior management?

- Are there any risks transferred? Is a corresponding insurance policy applicable?
- Are the risks associated with an intelligent system integrated into corporate and information security risk management?
- Is a business continuity plan established and approved, and does it include intelligent systems?

Metrics and Visibility:

- Are proper measurable metrics, such as KPIs or KRIs, in place to maintain transparency, and are they utilized effectively?
- Is there an AI Bill of Materials (AIBOM) in place to track all AI components?
- Is there a Software Bill of Materials (SBOM) in place to track all digital components of the AI?
- Is there a Bill of Materials (BOM) in place to track all physical components of the AI?
- Are assets within the AI system easily visible, and can they be tracked both physically and digitally?

Explainability AI Principles (XAI):

- Is there a clear understanding of the principles proposed for XAI (explanation, meaningful, explanation accuracy, and knowledge limits), and are these principles being applied appropriately?
- Are there mechanisms in place to explain the intelligent system's decision-making process to stakeholders?
- Is the explanation of outputs meaningful and understandable to the intended end-user?
- Does the system ensure explanation accuracy, reflecting the reason for generating the output correctly?
- Does the system operate only under conditions for which it was designed and/or when it reaches sufficient confidence in its output (knowledge limits)?
- How is the organization addressing the potential challenge of intelligent system's decisions becoming beyond human comprehension, especially with advanced systems?

Fairness and Bias:

- Is there a diverse dataset representing multiple demographic groups, socioeconomic backgrounds, and varied viewpoints that the intelligent system can use to ensure fairness in its training?
- Has the intelligent system been tested rigorously to ensure that it is fit-for-use in the intended context, and that bias has been minimized or eliminated?
- Is ongoing bias testing conducted to account for different environments or scenarios?

AI Resilience:

- Is the system resilient and tested under changed parameters and/or in unfamiliar environments, considering the criticality of the intelligent system?
- If the system incorporates plasticity (self-healing, self/continuous learning, evolutionary algorithms) approaches, is the plasticity transparent?
- If the intelligent system has plasticity, is it able to repurpose itself or change its limiting parameters? What is the design to prevent a drift from its original purpose or use case?

Privacy and Security:

- How well are safeguards in place for privacy and security?
- Does the system have a privacy-by-design and security-by-design approach throughout its lifecycle?
- Have the code/libraries been verified for security vulnerabilities?
- Were secure development strategies applied during the development phases?

Audit Process:

- How is critical thinking applied in the auditing process, especially considering the fast-evolving nature of intelligent systems?
- Are previous audit results available? Have identified findings been successfully remediated?

Appendix 3: Audit Questions "Accountability, Responsibility, and Liability"

Awareness of Accountability:

- Are senior management and other personnel aware of their accountability, responsibility, and liability?
- Is there proper training given to employees on accountability, responsibility, and liability to encourage responsible behavior and minimize legal risks?
- Is there proper documentation or evidence that outlines the assigned responsibilities and obligations of team members, and have these been clearly communicated to them?
- Are there any concerns related to the legal framework for holding individuals or organizations accountable for their responsibilities and the consequences of their actions?
- Are there mechanisms in place to mitigate any potential legal risks to the senior management?

- Are there plans to address risks associated with accountability, responsibility, and liability in AI insurance, and are these plans being communicated?

Reporting and Risk Management:

- Are there proper reporting and justification procedures in place when it comes to accountability, so that individuals or groups can provide explanations for their actions or decisions?
- Is a change management process established to track any changes reliably and transparently?
- Are there mechanisms in place to measure and track outcomes related to specific actions and decisions, and are individuals or groups held accountable for success or failure?
- How does the company handle the fact that success influences risk-appetite?
- How diverse is senior management?
- Are there potential risks associated with responsibilities being shared among team members, and if so, are these risks being mitigated?
- How does the company educate employees and management on responsible use of AI, including an understanding of the risks for AI hallucinations and guardrails on the ability to rely on the outputs?

Appendix 4: Audit Questions "Laws, Regulations, and Standards"

Legal and Regulatory Compliance:

- What laws and regulations apply to the intelligent system?
- Has the intelligent system been developed in compliance with these laws and regulations?
- How does the development team ensure ongoing compliance with any changes in laws, regulations, and standards?
- Are users of the intelligent system informed about their rights and the application's compliance with relevant laws, regulations, and standards?
- Are industry-specific standards and best practices (e.g., ISO/IEC 42001:2023, ISO/IEC 27701) in place and audited regularly?
- What data protection and privacy laws apply to the intelligent system?
- How is compliance with these privacy laws implemented or organized?
- Is a process established to communicate data breaches according to current law and within required timelines?

Privacy Risk Management:

- Are privacy risk factors assessed, documented, communicated, mitigated, accepted, or transferred (insurance)?
- Are measures taken to manage the withdrawal of consent to use a data owner's data or is the data owner informed accordingly?
- Can data owners limit how their data is used?
- Is there a record of explicit agreement to use personal data for further use and/or training of the intelligent system?
- Is the user aware that such an agreement possibly can't be revoked for data already provided to an intelligent system?
- Can data owners tag data according to their criticality, and does the intelligent system have the ability to treat data with different criticality according to the associated risks?
- How is data theft prevented and/or detected?
- How is a data breach handled? Is the remaining risk accepted or transferred?
- Is data disposed of securely?

Ethical and Sustainability Guidelines and Standards:

- Are there any ethical guidelines or standards that apply to intelligent systems?
- Is the implementation of ethical guidelines fully documented?
- Are there any sustainability guidelines or standards that apply to AI?
- How are sustainability aspects integrated into the intelligent system?

Governance and Accountability:

- Does the company have a list of all applicable laws, regulations, standards, and internal policies?
- Does the company have an inventory of all intelligent systems used within the company?
- Does the company have an overview of the third-party suppliers of their intelligent systems?
- Are auditable trails for data usage, model training, and decision-making processes recorded to ensure accountability?
- Has the intelligent system been subjected to any third-party audits or certifications to verify its compliance with relevant laws, regulations, and standards?

Fairness and Bias:

- Has the intelligent system been tested for fairness, bias, and discrimination against protected groups?
- How were the findings documented and, if necessary, addressed?

Appendix 5: Audit Questions "AI Business Case"

Business Case and Process:

- Does the intelligent system meet a specific business case?
- How was the business case for the AI system defined, and what are the expected metrics for its performance, resilience, and explainability?
- Are the business process(es) defined, including inputs and outputs?

Metrics and KPIs:

- Are measurable KPIs defined for performance, intelligent system resilience, and explainability (if required)?
- How do the expected metrics for the AI system's "fitness" align with its intended use and context?
- Are KPIs visualized, put in context, and reported to the CAIO or equivalent?
- Is a process in place to take action if KPIs are outside the acceptable range?

"By Design" Approach:

- Did the development of the intelligent system follow a proper "by design" approach, taking into consideration ethical principles, privacy, and security?

Audit Scope Definition:

- What is the scope of the audit, including the boundaries of the system, key stakeholders, relevant regulations, and specific concerns such as end-user safety, product quality, and data confidentiality, integrity, and availability (CIA)?
- Have all relevant stakeholders been identified and included in the audit?

Risk Assessment:

- How does corporate risk management integrate with the risk assessment for the AI system, and what processes are in place to ensure trustworthy AI?
- Are there any additional risks or concerns that should be considered when auditing the AI system, such as business continuity?

Senior Management Involvement:

- Is senior management fully aware and accountable for the deployed intelligent systems?

Critical Thinking and Assumption Challenge:

- How is critical thinking applied to connect seemingly independent dots and enhance the security of intelligent systems?
- How does the auditor apply outside-the-box thinking to challenge assumptions and mitigate potential risks?
- How does the auditor ensure that the AI system is "fit for use" in its specific environment and does not pose any long-term security risks?

Levels of Human Oversight:

- Has the organization clearly defined the required level of human oversight for the AI system (human-in-the-loop, human-out-of-the-loop, human-on-the-loop)?
- Is there documentation explaining the rationale behind the chosen level of human oversight?

Business Case Considerations:

- Does the initial business case include an assessment of the appropriate level of human involvement to mitigate risk or harm?
- How frequently is the level of human involvement reviewed and updated in the business case?

Human Review Options:

- Are "opt-out" and/or "opt-in" options for human review implemented within the AI system?
- Is there a mechanism that allows a human to challenge the decisions made by the intelligent system?

Effectiveness and Monitoring:

- Is there a process in place to monitor the effectiveness of human oversight and adjust it as necessary?
- Are there training programs for humans involved in the oversight process to ensure they understand their roles and responsibilities?

Risk Assessment:

- Has a risk assessment been conducted to determine the potential harm from inadequate human involvement in the decision-making process of the AI system?
- How does the organization ensure that the chosen level of human oversight aligns with the criticality of the system and potential impact on stakeholders?
- How are stakeholders informed about their ability to opt-in or opt-out of human review and their right to challenge intelligent system decisions?

Compliance and Best Practices:

- Does the organization follow industry best practices and regulatory requirements when deciding on the level of human oversight and review options?
- How does the organization ensure compliance with these practices and regulations over time in regards to human oversight?

Appendix 6: Audit Questions "AI Infrastructure"

Infrastructure Qualification:

- What qualifications does the existing infrastructure have to handle AI-specific workloads?
- What were the processes and criteria used for infrastructure qualification?
- Has the infrastructure been audited for security, privacy, scalability, and sustainability?
- Are there any relevant regulations or standards with which the infrastructure needs to comply, and has compliance been checked during the audit process?
- Has the auditor thought beyond current regulations and ensured that the audit will help with long-term trustworthiness of the intelligent systems?
- What disaster recovery plans are in place to handle potential infrastructure failures?

- What is the (geopolitical and physical) security of data centers used in the infrastructure?

Infrastructure Performance and Scalability:

- Are there any current limitations or challenges in the infrastructure that could impact the performance or reliability of the intelligent system through factors like computational power, data processing speeds, and bandwidth?
- How scalable is the infrastructure in response to increasing data volumes and computational needs?
- Explain the capacity planning strategies and technologies used to facilitate scalability.

Data Handling and Privacy in Infrastructure:

- How does the intelligent system infrastructure handle sensitive data?
- Is confidential computing integrated into the hardware to ensure advanced privacy?
- What measures are in place to secure data during processing and transfer between different parts of the infrastructure?

Sustainability and Environmental Impact:

- How does the intelligent system infrastructure address energy efficiency, such as the use of energy-efficient hardware?
- Outline the energy-saving features and practices adopted in the infrastructure design and operation.
- Are renewable energy sources being used?
- How does the infrastructure address environmental sustainability?
- How is electronic waste disposed of?

Appendix 7: Audit Questions "Sensors"

Sensor Integration and Interoperability:

- Are all necessary sensors, including visual, sound, touch, temperature, proximity, and chemical compound detectors, properly incorporated into the intelligent system?
- Are these sensors delivering the expected data?
- What are the protocols for sensors to communicate and share contextual data?

- For context-aware sensor systems, are sensor technology, data analytics, networking protocols and machine learning (including TinyML) being utilized appropriately?

Risk Management and Compliance:

- Are there any underestimated risks associated with the combined capabilities of sensors?
- Is there a plan in place to mitigate any potential problems of neglect, misuse, or intentional abuse of such sensors?
- Are proper analytical tools in place to evaluate the trustworthiness of sensor data in generating comprehensive insights for intelligent systems?
- Are proper data protection measures in place to ensure the privacy and security of the data obtained by the sensors?
- Is the use of sensors in intelligent systems compliant with all applicable laws?

Technological Frontiers:

- Are there any emerging technologies such as miniaturization, context-awareness, and sensor fusion that could influence the overall risk assessment of the intelligent system?
- When replacing sensors, are expected benefits such as improved accuracy, enhanced automation, and reduced power consumption being achieved?

Appendix 8: Audit Questions "Data"

Data Sources and Requirements:

- Who defines the data requirements (quantity, quality, features) for the intelligent system's business case?
- What metrics and methods are used to assess data quality (accuracy, consistency, completeness, reliability, relevance)?
- What measures are in place to ensure that the data is of high quality?
- How do the data characteristics and/or quality impact the performance of the system?
- What kind of data was used for the specific business case (e.g., labeled, unlabeled, semi-labeled, structured, unstructured, organic only, synthetic only, hybrid approach)?
- How are these data requirements documented and communicated?
- What justification exists for using external/purchased data?
- How is data sourced, collected, and curated for the intelligent system?

- Is data labeled regarding its source and whether it is of organic or synthetic origin?
- How is the risk of unwanted bias in data mitigated?

Data Provenance and Scraping:

- How is data provenance tracked to ensure traceability from collection to usage (“black market data,” "dirty data")?
- How are the risks associated with the unintentional use of "black market data" mitigated?
- What procedures are in place to identify and avoid "dirty data"?
- What mechanisms are in place to verify the source and lineage of all datasets used?
- Is training data origin traceable ("data pedigree") and verified or certified by a trusted organization?
- Are scraped sources documented to ensure traceability and data origin?
- Does data scraping respect the Terms of Service (TOS) of websites to prevent violations?
- How is compliance with copyright regulations ensured for training datasets, particularly for generative AI?
- Do mechanisms ensure that generative AI systems do not replicate original (art)work, which is protected by copyright regulations, patents, or similar?
- What internal processes exist for identifying and addressing potential copyright infringement?
- Are agreements for copyright-protected work documented?

Data Management and Lifecycle:

- What procedures are in place to ensure privacy and compliance with the CIA triad (confidentiality, integrity, availability) during data collection?
- What documented policies and procedures govern the entire data lifecycle (creation/acquisition, use, transfer, storage, deletion)?
- Is there an audit trail of data accessed, edited, or deleted by the intelligent system and can this be traced back to the user who authorized such changes?
- How do changes or updates in the data used by the intelligent system affect its performance?
- What measures are in place to ensure that the system remains accurate and effective in its decision-making process during its whole operational life (and beyond)?

Data Privacy and Security:

- How is data de-identification (anonymization/pseudonymization) implemented for privacy protection?
- How is compliance with differential privacy techniques ensured when applicable?
- Are measures implemented to blur/encrypt personal data before it is submitted into a training dataset?
- How is data retention regulated and achieved?
- What measures are used to ensure data security at rest and in transit?
- Are post-quantum cryptography solutions considered for future-proof encryption?

Data Preparation Processes:

- Who is responsible for data preparation and splitting (training, validation, and testing)?
- How is data splitting documented and justified?
- What procedures are in place to ensure data integrity during preparation?

Synthetic Data:

- How is synthetic data generation validated for suitability in the model of the AI system?
- How is the generation of synthetic data regulated and documented?

Data Quality Assurance:

- Who is responsible for data quality assurance and adherence to data governance policies?
- How are findings from data quality assessments documented and addressed?

Data Governance Framework:

- Is there a well-defined data governance framework in place?
- Are there policies and procedures to ensure ethical standards are maintained in data selection, manipulations, and usage?

Model Output:

- How is model output data stored and secured?

Appendix 9: Audit Questions "Data Processing Unit"

Data Processing Units:

- Are all necessary data processing units available, including CPUs, GPUs, TPUs, and edge processors?
- Is each data processing unit being utilized efficiently and effectively for the specific tasks for which it is designed?
- Are there any limitations or challenges with the existing data processing units when it comes to the scale of the data processing needs?
- Is there a proper mechanism for integrating different data processing units and for allocating workloads so that tasks are appropriately distributed?
- Are edge processors being utilized to enhance efficiency by processing data locally, reducing latency in IoT environments?

Security and Compliance:

- Are there proper security measures in place to protect data processing units from cyberattacks?
- Is there proper documentation related to data processing units that includes details such as their technical specifications and integration into the system?
- Are there any policies or guidelines in place that govern the usage of different data processing units and their associated technologies in compliance with regulatory requirements?

Emerging Technologies:

- Are there any emerging technologies like quantum computers being evaluated for increasing computational power, and if so, what measures are in place to manage their potential integration with the existing infrastructure?
- Are biological computers being explored for processing, and if so, are there plans to explore their use in enhancing intelligent system capabilities?

Appendix 10: Audit Questions "Data Storage"

Data Storage Solutions:

- Are all necessary storage solutions available, including internal storage options (SSDs and HDDs), RAM, and cloud storage solutions?
- Are the internal storage options properly integrated into the system, and is there enough storage to meet its requirements?
- Are RAM resources configured to provide adequate data retrieval for immediate processing needs?
- Is the cloud storage solution deployed correctly and providing scalable and accessible repositories for storing and analyzing large datasets?
- Are there any concerns about data migration and integration from internal storage options to cloud storage solutions?
- Is there a plan or process in place to mitigate any potential data migration problems?
- Is there proper data backup and recovery capability to ensure data availability and integrity even in the event of a system crash?

Data Security, Compliance, and Documentation:

- Are there any data security concerns, and is data stored in any of the storage solutions encrypted or adequately protected by access control mechanisms?
- Is there proper documentation on how to use each storage option, how to integrate storage options properly, and how to backup and recover data?
- Are there considerations for regulations such as GDPR, ensuring compliance in storing and managing data, and are necessary measures in place?

Emerging Storage Technologies:

- Are there plans to integrate advanced storage methods like holographic data storage and DNA-based storage into the intelligent system?

Appendix 11: Audit Questions "Networking, Connectivity, and Communication Interfaces"

Integration and Configuration of Communication Interfaces and Protocols:

- Are the necessary communication interfaces properly integrated into the intelligent system to ensure seamless data exchange?
- Are there proper guidelines and procedures in place for selecting the most appropriate communication interfaces based on usage scenarios and device compatibility?
- Are traditional networking methods, such as WLAN, Ethernet, and cellular (4G/5G), configured appropriately to maintain optimal network performance?
- Are short-range communication protocols, such as Bluetooth, BLE, and RFID, adequately secured against wireless vulnerabilities and data breaches?
- Are IoT-specific protocols, like LoRaWAN, Zigbee, Z-Wave, NB-IoT, and Sigfox, being used for low-power, long-range, and short-range communications, and is adequate security in place to protect data transmitted over these protocols?
- Are industrial and real-time communication protocols like OPC-UA, DDS, and CoAP being utilized to meet machine-to-machine communication and high-performance data exchange needs, and is data secure?

Interoperability and Cybersecurity:

- Are there any concerns about interoperability between different communication protocols used in these intelligent systems, and is there a plan in place to mitigate any potential problems?
- Is there a plan in place to manage cybersecurity challenges related to the various communication protocols, from wireless vulnerabilities to securing data transmitted over constrained networks?

Future Connectivity Plans:

- Are there plans to upgrade to 6G networks to improve connectivity and reduce latency, and if so, are there steps in place to ensure a smooth transition?
- Are there plans to implement emerging connectivity technology, such as quantum communication networks?

Appendix 12: Audit Questions "Fog and Cloud Computing"

Fog Computing Infrastructure:

- Where are the fog computing nodes strategically placed to minimize latency and maximize efficiency in the data transmission process?
- What level of processing power and storage capability do the fog computing nodes possess? Is this enough to effectively perform local data analysis, filtering, and pre-processing?
- Where are intelligent gateways positioned on the fog computing nodes to help optimize the overall cloud load and ensure real-time application performance?
- Are fog computing systems designed to integrate seamlessly with cloud computing systems, allowing for a smooth flow of data between the two systems?
- Are there any current or potential interoperability issues between different fog computing implementations that could affect the integration of different fog computing systems?
- Are there enough resources to implement intelligent systems for predictive analytics and edge AI applications at the fog level, and are such applications expected to improve decision-making capabilities?
- Are there any concerns about the scalability of fog computing solutions, especially when it comes to supporting emerging technologies such as autonomous vehicles, smart cities, and industrial IoT?

Security and Compliance:

- What cybersecurity measures are currently in place to ensure the security of data transmitted over fog computing networks?
- Are there any policies or guidelines in place to ensure that fog computing solutions meet the necessary regulatory requirements for data intake, storage, and transmission?
- Are there any backup systems or plans in place in case of internet disruptions that could affect the operation of fog computing networks?

Appendix 13: Audit Questions "Software Components"

Installation and Configuration:

- Who is responsible for ensuring that all necessary software components (virtualization software, operating systems, middleware, application software, data management tools, monitoring, and logging tools, and container orchestration platforms) are properly installed and integrated?
- How are the virtual machines properly configured and allocated with the necessary resources based on their demands?
- Is the middleware capable of providing seamless communication between different applications and systems?
- Is the container orchestration platform efficiently managing and scaling AI systems in a distributed environment?
- Are there any limitations or challenges in using software components for scalability, reliability, and efficiency?

Maintenance and Documentation:

- How often are the software components updated and security vulnerabilities patched?
- Where is the proper documentation related to the installation, management, and maintenance procedures of the software components located?
- Are monitoring and logging tools equipped with the necessary features and functionality to track system performance and security adequately?

Appendix 14: Audit Questions "Algorithms, Training, and Models"

Algorithms:

- How is the choice of algorithms justified based on the problem analysis (e.g., learning type, data availability)?
- Are there potential biases associated with the chosen algorithms?
- If Explainable AI (XAI) techniques are relevant, are they implemented to understand the model's reasoning?

Training Datasets:

- How is the quality and representativeness of the training data ensured?
- What measures are in place to mitigate bias and fairness concerns within the datasets?
- Are there procedures for identifying and addressing data containing privacy violations (e.g., personally identifiable information)?
- How is the balance between labeled, unlabeled, and synthetic data managed for effective training?

Pre-Trained Models:

- If pre-trained models are used, how is their provenance and potential biases documented and understood?
- Are there licensing restrictions or copyright considerations associated with the pre-trained models?

Model Training Frameworks:

- Are the chosen frameworks (e.g., TensorFlow, PyTorch) secure and up to date?
- How are version control and reproducibility practices implemented for model training?

Evaluation Metrics:

- How is a diverse set of evaluation metrics (e.g., accuracy, precision, recall) used to assess model performance beyond just basic accuracy?
- Are the limitations of chosen metrics understood and considered when interpreting results?

Hyperparameter Tuning:

- How is the process of hyperparameter tuning documented and validated to avoid overfitting?
- Are there procedures to prevent human bias from influencing hyperparameter selection?

Model Inference Engines:

- What security measures are in place to protect model inference engines from adversarial attacks or manipulation?
- How are the efficiency and performance of inference engines monitored and optimized for real-world deployment?

Model Deployment and Management:

- How is the model deployed in a production environment to ensure continuous monitoring and performance evaluation?
- Are rollback strategies established in case of model drift or performance degradation?

Security and Drift Detection:

- How are security vulnerabilities in the model and system assessed and addressed?
- Are there processes for detecting and mitigating adversarial attacks (e.g., poisoning, evasion)?
- How is data drift (concept drift) monitored and mitigated to maintain model effectiveness over time?

Ethical Considerations:

- How are fairness, bias, and explainability considerations integrated throughout the intelligent system lifecycle?
- Are there procedures for identifying and mitigating potential ethical risks associated with the model's outputs?
- If the model is used in high-stakes domains (e.g., healthcare, finance), are there additional ethical and regulatory considerations addressed?

Transparency and Explainability:

- Is there a clear understanding of the model's limitations and potential failure modes?
- Are there mechanisms for explaining the model's decisions to human users, especially in critical scenarios?
- How is human oversight maintained throughout the development, deployment, and operation of the intelligent system?

Appendix 15: Audit Questions "Fine-Tuning and Validation"

Hyperparameter Tuning:

- Were the hyperparameters tuned appropriately before training the model?
- What was the rationale behind the chosen values?
- Were any search or optimization strategies used?

Fine-Tuning the Model:

- Were attempts made to improve the accuracy of the model by fine-tuning it?
- What approach was used for fine-tuning?
- Were additional data sources or different methods considered for fine-tuning?
- Were any search or optimization strategies used?

Validation Methodology:

- What was the chosen validation methodology, and how was the validation dataset selected?
- How was the performance of the model evaluated, and what evaluation metrics were used?

Overfitting Checks:

- Was a test for overfitting performed, and if so, how?
- Was a separate test dataset used, or did the model use a validation dataset for this purpose?

Performance and Robustness:

- How was the accuracy and reliability of model outputs verified?
- What procedures are in place for continuous monitoring of model performance and potential biases?

Generalization and Robustness:

- What was done to check the model's robustness and generalization capabilities?
- Were any adversarial attacks or other perturbations used to test the robustness of the model?
- Was pentesting performed to evaluate the security of the AI?
- Is a bug bounty set up to continuously check for vulnerabilities?

Interpretability and Transparency:

- What measures were taken to ensure the interpretability and transparency of the model?
- Were any mitigating techniques applied to prevent bias?

Performance Metrics and Criteria:

- How was the performance of the AI system's model measured against the requirements?
- What were the acceptance criteria for the model's deployment?
- Who made the final decision to deploy the model, and on what basis?

Assumptions and Limitations:

- What are the assumptions and limitations of the AI system's model, and how are they documented?

"Fit-For-Use" Validation:

- Has the AI system's model been validated for its intended function, in realistic scenarios, and under a range of possible environmental conditions?
- Does the AI system's model produce results that are consistent with expectations, based on the nature of the data and the intended application of the model?

Performance and Reliability:

- Is the AI model's performance acceptable in terms of accuracy, speed, reliability, and resource requirements?
- Has the AI system's model's behavior been adequately characterized under nominal and adverse conditions, including edge cases, outliers, and unexpected inputs?

Bias and Unintended Effects:

- Does the AI system's model's output make sense from a semantic, contextual, or human-interpretable perspective?
- Are there any biases or unintended effects that might result from the use of the AI system's model in practice?

Independent Evaluation:

- Has an independent third-party evaluation of the AI model's performance and validation been performed?
- What were the evaluation criteria used in the third-party evaluation?

Appendix 16: Audit Questions "Actuators"

Actuator Choice Justification:

- What criteria were used to select the specific types of actuators (e.g., motors, servos, grippers)?
- How do the chosen actuators align with the intended force, speed, precision, and range of motion requirements of the robot?
- Were alternative actuator types considered and evaluated?

Documentation and Testing:

- Is there comprehensive documentation detailing the specifications, capabilities, and limitations of each actuator used?
- What pre-deployment testing procedures were conducted to ensure actuator reliability and performance under expected operational conditions?

Actuator Performance and Reliability:

- What metrics are used to measure the performance of actuators during operation (e.g., force output, speed, precision)?
- How frequently are these performance metrics evaluated, and what are the thresholds for acceptable performance?

Durability and Maintenance:

- What are the expected lifespans of the various actuators, and how is durability assessed?
- What maintenance schedules and procedures are in place to ensure ongoing actuator performance and prevent failures?

Fault Detection and Management:

- How are actuator faults detected, reported, and managed?
- What redundancy measures are implemented to ensure continued operation in the event of actuator failure?

Safety and Risk Management:

- What safety mechanisms are in place to prevent unintended movements or actions by the actuators?
- How are potential risks from actuator malfunctions or misuse mitigated?

Emergency Protocols:

- What emergency stop protocols and safety checks are integrated into the actuator control systems?
- How are these protocols tested and validated?

Ethical Considerations:

- How are ethical considerations, particularly around human-like movements and interactions, addressed in the design and deployment of actuators?
- What measures are taken to prevent misuse of actuators in ways that could cause harm or ethical concerns?

Innovative Actuators:

- Are there any advanced or experimental actuators being used (e.g., artificial muscles, microfluidic technologies)?
- What potential benefits and risks are associated with these new technologies, and how are they managed?

Adaptability and Scalability:

- How adaptable are the actuators to future advancements or changes in the robot's operational requirements?
- Are there provisions for upgrading or scaling the actuator systems without significant redesign?

Compliance and Beyond:

- How do the actuators comply with relevant industry standards and regulations?
- What internal and external audits are conducted to ensure ongoing compliance?
- Has a third-party evaluation of the trustworthiness posture of the AI been performed? What was the result?
- What innovative practices or technologies are being employed to go beyond standard compliance, enhancing actuator safety, reliability, and efficiency?
- How does the organization stay ahead of emerging trends and potential future regulatory changes in actuator technology?

Environmental and Operational Conditions:

- How are actuators tested and validated for operation in various environmental conditions (e.g., temperature, humidity or water, dust, UV exposure, limited visibility, extreme noise or vibration)?
- What measures are in place to protect actuators from environmental damage or degradation?
- How can the system be disposed of in an environmentally acceptable way?

Operational Versatility:

- How versatile are the actuators in handling different operational tasks or adapting to new tasks?
- What training or calibration processes are in place to optimize actuator performance for specific applications?
- Is business continuity possible without certain actuators? Is there a business continuity plan in case of production bottlenecks or material shortages?

Appendix 17: Audit Questions "Power Supplies"

Compliance with Standards:

- How do the power supplies comply with relevant industry standards and regulations?
- What internal and external audits are conducted to ensure ongoing compliance?

Beyond Compliance:

- What innovative practices or technologies are being employed to go beyond standard compliance, enhancing power supply safety, reliability, and efficiency?
- How does the organization stay ahead of emerging trends and potential future regulatory changes in power supply technology?

Power Supply Choice Justification:

- What criteria were used to select the power supply options (e.g., batteries, AC power, DC power, hybrid systems) for intelligent systems?
- How do these power supply options align with the operational requirements and constraints of the AI system (e.g., mobility, runtime, energy efficiency)?

Documentation and Testing:

- Is there comprehensive documentation detailing the specifications, capabilities, and limitations of each power supply used?
- What pre-deployment testing procedures were conducted to ensure power supply reliability and performance under expected operational conditions?

Performance:

- What metrics are used to measure the performance and efficiency of power supplies during operation (e.g., battery life, energy consumption, charging time)?
- How frequently are these performance metrics evaluated, and what are the thresholds for acceptable performance?
- How is the performance of power supplies continuously monitored, and what tools are used for this purpose?
- What actions are taken when power supply performance metrics indicate potential issues?

Durability and Maintenance:

- What are the expected lifespans of the various power supplies, and how is durability assessed?
- What maintenance schedules and procedures are in place to ensure ongoing power supply performance and prevent failures?

Fault Detection and Management:

- How are power supply faults detected, reported, and managed?
- What redundancy measures are implemented to ensure continuous operation in the event of power supply failure?

Power Management Systems:

- How are power management systems optimized to ensure efficient power consumption across intelligent systems?
- What measures are in place to dynamically manage power usage based on operational needs and energy availability?

Safety Mechanisms:

- What safety mechanisms are in place to prevent hazards associated with power supplies, such as overheating, overcharging, or electrical surges?
- How are potential risks from power supply malfunctions or misuse mitigated?

Emergency Protocols:

- What emergency shutdown protocols and safety checks are integrated into the power management systems?
- How are these protocols tested and validated?

Wireless Charging:

- How is wireless charging technology integrated and tested for intelligent systems, particularly robots?
- What measures are in place to ensure the system can shut off in case of emergency or malfunction during wireless charging?

Environmental Considerations:

- How is the environmental impact of power supply choices assessed, particularly regarding battery disposal and energy consumption at data centers?
- What steps are taken to minimize the carbon footprint and enhance the sustainability of intelligent system power supplies?

Renewable Energy Integration:

- To what extent are renewable energy sources used to power intelligent systems, particularly in data centers?
- What strategies are in place to increase the use of renewable energy and reduce reliance on non-renewable sources?

Innovative Power Solutions:

- Are there any advanced or experimental power supply technologies being used?
- What potential benefits and risks are associated with these new technologies, and how are they managed?

Adaptability and Scalability:

- How adaptable are the power supplies to future advancements or changes in the operational requirements of the AI system?
- Are there provisions for upgrading or scaling the power supply systems without significant redesign?

Appendix 18: Audit Questions "User Interfaces"

Display Quality and Durability:

- How is the quality and durability of displays/screens assessed to ensure reliable user interaction?
- Are there measures in place to prevent screen glare, distortion, or visibility issues under varying environmental conditions?
- Is there the possibility to enhance privacy by applying a special filter physically or digitally?

Touchscreen Functionality:

- How are touchscreens tested for responsiveness and accuracy across different touch gestures (e.g., tapping, swiping)?
- Are there protocols for calibrating touchscreens to maintain optimal performance over time?
- Is the sensitivity of the touchscreen adaptable to varying temperatures or handling with gloves?

Voice Recognition Accuracy:

- How are accuracy and reliability measured, considering accents and background noise?
- Are there mechanisms to update voice recognition algorithms based on user feedback?

Noise Reduction and Enhancement:

- Is noise cancellation available to improve recognition in noisy environments?
- Does the system offer speech enhancement features to improve clarity?
- Can specific frequencies be blocked to reduce unwanted noise?

Environmental Factors:

- How does vibration, such as from phones being in pockets, affect recognition?
- How well does the system handle interference from other voices or sounds?

Control Panel Usability:

- How are physical control panels designed to optimize user interaction with buttons, knobs, and levers?

- Are ergonomic principles considered to ensure ease of use and accessibility for all users?
- Are safety and security principles considered to ensure the possibility to use the panel under varying conditions like extreme temperature, low visibility, high stress, high acceleration, or increased gravitation?

Mobile App Security:

- What security measures are implemented to protect user data and prevent unauthorized access through mobile applications?
- How are mobile apps tested for compatibility across different devices and operating systems?
- Is continuous and end-user-independent wireless security patching enabled?
- Are wireless firmware updates possible considering cybersecurity threats?
- Is an emergency shutdown enabled in case the app malfunctions or doesn't respond?
- Does the app have self-diagnostic abilities to check for proper functioning? Is it enabled to take action if malfunctioning is detected?
- Does the app have self-healing abilities? How are they controlled?

Gesture Recognition Accuracy:

- How accurate and reliable is the gesture recognition system in interpreting hand or body movements?
- Is continuous learning enabled to train the system to improve the recognition of personalized gestures?
- Are privacy measures in place to secure identifiable gestures or movement patterns?
- Are adjustments made to optimize recognition capabilities for varying lighting conditions and user gestures?
- Is there a safety protocol implemented to verify gestures that trigger specific actions?

AR/VR Integration and Emerging Technologies:

- How seamlessly is AR/VR technology integrated into the user interface, and what measures ensure immersive user experiences?
- Are there protocols in place to minimize motion sickness and optimize visual fidelity in VR environments?
- What environmental security measures are implemented to ensure the safe operation of AR/VR systems?

- How is it assessed that users do not confuse AR/VR experiences with reality in a manner resembling psychosis?
- What measures are in place to prevent unintentional self-harm in AR/VR environments?
- How is unintentional harm to others prevented in AR/VR environments?
- What technologies or methodologies are implemented to differentiate between AR/VR simulations and reality?

Security:

- How is physical access to the BCI system controlled to prevent unauthorized tampering or damage?
- How is the data privacy of neural data ensured during collection, transmission, and storage?
- How are physical and cybersecurity access controlled to prevent unauthorized interaction and data breaches?
- Are protocols in place for secure storage, transmission, and anonymization of user data collected by haptic feedback systems?

Malfunction and Safety:

- What protocols are in place to detect and address malfunctions in the BCI hardware or software?
- Are there safeguards to prevent unintended actions or responses triggered by neural signals?
- How are users informed about potential BCI risks and safety precautions?

Accountability and Responsibility:

- Who bears responsibility for actions performed by BCI-controlled devices?
- Are there protocols for determining accountability in case of BCI misuse or malfunctions?
- How is user consent and understanding of potential risks documented?

Insurance and Liability:

- Is there insurance coverage or liability protection for BCI users and stakeholders?
- How are potential liabilities assessed and mitigated in case of BCI-related accidents or failures?

Ethical Considerations:

- What ethical guidelines govern BCI development and deployment (informed consent, autonomy, equity)?
- How are potential biases in BCI algorithms or data collection addressed?
- Do the user interface design documents and standards adequately incorporate accessibility and inclusivity requirements?
- Has usability testing been conducted with users with disabilities to confirm that the user interface is accessible and usable?
- Has the user interface code been reviewed to ensure compatibility with assistive technologies and adherence to accessibility guidelines?
- Have linguistic and cultural reviews been conducted to ensure the user interface is inclusive of diverse languages and cultures?
- Has the user interface design been reviewed to ensure it is free from cultural biases and stereotypes?
- What measures prevent discomfort, injury, or sensory overload from haptic feedback?
- How are ethical considerations like informed consent, user autonomy, fairness, and inclusivity addressed in haptic feedback applications, especially in sensitive domains?

Performance, User Experience and Feedback Loop:

- How is the user experience optimized for comfort, ease of use, and effective BCI interaction?
- Are there provisions for adapting the BCI to individual user needs (customization, user feedback)?
- How does the organization incorporate user feedback into continuous UI/UX improvement cycles?
- How are accuracy, consistency, and calibration of haptic feedback ensured across users and environments?
- How is the user experience optimized for intuitive interaction, with mechanisms to adjust haptic parameters based on user preferences and sensory needs?
- Are there mechanisms for real-time monitoring of user satisfaction and usability metrics to inform iterative design updates?
- How are accuracy, consistency, and calibration of haptic feedback ensured across users and environments?

Regulatory Compliance and Innovation:

- How does the BCI comply with relevant standards for medical devices or consumer electronics?
- Are there ongoing efforts to monitor and update compliance with evolving regulations and best practices?
- What processes are in place to monitor and maintain compliance with evolving regulations?
- How does the organization gather user feedback and foster innovation in haptic feedback technology while prioritizing user safety and reliability?

Risk Assessment and Mitigation

- How are potential risks associated with UI/UX technologies identified and mitigated proactively?
- Is there a framework for conducting scenario-based risk assessments to prepare for unforeseen challenges or vulnerabilities?

Research and Development (R&D) and Future Technology Integration:

- What research initiatives are in place to improve BCI performance and address emerging challenges?
- How does the organization foster innovation while ensuring responsible BCI development and deployment?
- How is the system prepared to integrate upcoming UI/UX technologies, such as advanced intelligent system interfaces or quantum computing-driven interactions?
- Are there strategies in place to monitor emerging risks and opportunities in UI/UX technology advancements?
- How seamlessly do haptic feedback systems integrate with other platforms? Are there interoperability standards to ensure compatibility with diverse devices and applications?

Appendix 19: Audit Questions "Control Interfaces"

Single Point of Failure:

- How is the risk of a single point of failure mitigated in centralized control units?
- Are there backup systems or redundancy measures in place to handle control unit failures?

Performance Monitoring:

- How is the performance of the centralized control unit monitored and maintained?
- Are there protocols for real-time diagnostics and failure prediction?
- Is there a failover procedure or redundancy implemented for the minimal required functionality of critical systems?

Security:

- What security measures are implemented to protect the centralized control unit from cyberattacks?
- How is access to the centralized control unit controlled and logged?
- Is a self-diagnostic ability implemented that is able to take action (e.g., emergency shutdown, maintaining essential performance, self-healing capabilities)?
- Can the control unit be shut down in case of an emergency?

Decision-Making Distribution:

- How is decision-making distributed across multiple units, and what protocols ensure consistency and coordination?
- Are there mechanisms to handle conflicts or contradictions between decentralized units?

Robustness and Adaptability:

- How is the system's robustness and adaptability to changing environments evaluated?
- Are there procedures for dynamically reallocating tasks among decentralized units based on real-time conditions?
- Can workloads be allocated dynamically depending on specific decentralized abilities leveraging different computational approaches (e.g., traditional computing using different processors, quantum computing, biological computing)?

Communication and Synchronization:

- How is communication between decentralized units managed and secured?
- Are synchronization protocols in place to ensure cohesive system behavior?
- Do the decentralized units manage to generate context awareness? Is this a benefit or a risk?

Goal Setting and Execution:

- How are high-level goals set by the central unit translated into specific actions by lower-level units?
- Are high-level goals checked by independent units (four-eyes-principle)?
- Are there feedback loops to inform the central unit about the status and performance of lower-level actions?

Autonomy and Coordination:

- How is the autonomy of high-level units controlled?
- How much autonomy do lower-level units have in hierarchical control systems?
- Are there mechanisms to ensure coordination and prevent conflicts between hierarchical levels?

Flexibility and Scalability:

- How flexible and scalable is the hierarchical control system for different applications and environments?
- How is scalability controlled?
- Are there protocols for reconfiguring the hierarchy based on system or user demands?

Behavior Definition and Activation:

- How are behaviors defined and prioritized in the system?
- How are behaviors controlled?
- What criteria are used to activate specific behaviors based on sensor inputs?
- What criteria are used to deactivate specific behaviors?

Efficiency and Complexity:

- How is the efficiency of behavior-based control systems assessed, especially for complex tasks?
- Are there measures to optimize the execution of pre-defined behaviors for better performance?
- How is the optimization of executive functions linked to increased bias, reduced fairness, or the neglect of context?
- Can energy consumption be optimized for sustainability without losing overall functionality?

Adaptability:

- How adaptable are the behavior-based control systems to new or unexpected situations?
- Are there mechanisms for learning and incorporating new behaviors based on experience?

Integration of Architectures:

- How are different control architectures integrated within the hybrid control system?
- Are there protocols to manage and leverage the strengths of each approach effectively?
- How is the balance between centralized planning and decentralized reactivity maintained?
- Are there adaptive strategies to shift the balance based on real-time requirements?

Future-Proofing:

- How future-proof is the hybrid control system in terms of scalability, adaptability, and technological advancements?
- Are there plans to integrate emerging technologies and control methodologies?

Safety and Reliability:

- How is the safety and reliability of the control system ensured across different operating conditions?
- Are there emergency protocols and fail-safes to handle unexpected failures or malfunctions?

Ethical and Legal Compliance:

- How is compliance with ethical standards and legal regulations ensured in the control system design and operation?
- Are there measures to address accountability and responsibility for actions taken by the control system?

Sustainability:

- How does the control system design consider environmental sustainability and energy efficiency?
- Are there initiatives to minimize the ecological footprint of the control system throughout its lifecycle?

Appendix 20: Audit Questions "Safety Systems"

Physical Safeguards and Mechanisms:

- How are physical safeguards such as mechanical barriers, enclosures, collision detection and avoidance, and limitations on speed or force designed to prevent physical harm?
- How are these safeguards tested and evaluated for effectiveness?
- How is a "kill switch" mechanism implemented, ensuring that it works as intended and is accessible under all circumstances? How is it tested and evaluated?

Sensor Fusion and Reliability:

- How is data from multiple sensors combined (i.e., sensor fusion) to create a more comprehensive and context-aware picture of the environment, aiding in safe navigation?
- Can sensor fusion impose different or novel risks, and how are those mitigated?
- How are sensors evaluated for their reliability and accuracy?

System Redundancy and Fault Tolerance:

- How are systems designed with redundancy in critical components to ensure continued safe function even if one component fails?
- How are these components evaluated for reliability?

Software and Decision-Making Integrity:

- How do software algorithms monitor system behavior and intervene to prevent unsafe actions?
- What measures are in place to ensure that the intervention decisions are accurate and reliable?
- Can the decision-making be challenged by a human (or another system)?
- How is ethical decision-making ensured, and what risk assessments are conducted to identify potential unforeseen consequences of physical actions or vulnerabilities arising from deeply interconnected systems?

Safety Protocols in Human-Robot Interaction:

- How are safety protocols and measures developed to prevent accidents?
- How is clear communication between humans and robots working together ensured?
- How are these protocols tested and evaluated? Are they continuously monitored?
- How is trustworthiness evaluated and enhanced on all levels, and what measures are taken to ensure that the intelligent systems are safe, secure, and reliable?

Appendix 21: Audit Questions "Security Systems"

Encryption and Key Management:

- How is data confidentiality ensured through encryption during both storage and transmission?
- How is key management implemented to ensure cryptographic keys are only accessed on a need-to-know basis, are refreshed/rotated according to policy, are created with the appropriate strength and ciphers, and are generated only by approved and trusted generators?
- Are the algorithms and protocols used following the recommended guidelines and standards, and are they regularly reviewed and updated to keep up with the increasing capabilities of attackers to break them?
- Have dual encryption techniques been implemented to prepare systems for resilience against future quantum computing threats?

Data Sanitization:

- How is the secure deletion of unnecessary or expired data ensured to prevent it from being recovered?

Authentication and Access Control:

- How is the legitimacy of people and system accounts requesting access confirmed?
- Which authentication methods are used to ensure that only authorized entities gain access?
- How is access to system resources regulated, and what predefined rules exist to assign permissions?
- How is unauthorized access to sensitive data and functionalities prevented?

Security Monitoring and Incident Management:

- How is security data from multiple sources integrated and analyzed to provide a centralized platform for threat detection, incident response, and comprehensive security monitoring?
- How is ongoing monitoring of the activities of systems and networks implemented to detect anomalies and potential security incidents in real-time and to enable proactive threat detection measures?
- How are security vulnerabilities actively identified and mitigated in software and hardware components? Through regular updates and patching?
- How is a structured approach developed and maintained to address and manage a security breach or cyberattack, aiming to limit damage, reduce recovery time and costs, and maintain business continuity?

Disaster Recovery and Business Continuity:

- Are backups performed and tested for effectiveness at regular intervals to guarantee the required recovery time and point objectives established by the risk assessment or as required by contracts or regulations?
- Are backups specifically protected against unauthorized access and attacks?
- How are plans and mechanisms for the recovery of critical data and systems implemented, in the event of significant disruption, such as a cyberattack, software or hardware failure, or natural disaster?

Security Audits and Physical Security:

- How are periodic security audits conducted to evaluate the effectiveness of security measures, ensure compliance with industry standards, and identify areas for improvement?
- How is the physical hardware of intelligent systems protected from theft, tampering, or damage?

Security Awareness and Training:

- How are users and operators of intelligent systems educated on potential security risks, best practices, and incident response procedures to maintain a trustworthy security posture?

Appendix 22: Audit Questions "Development and Debugging"

Development and Testing Platforms:

- How do collaborative development platforms enable code co-pilots, version control, and the ability to revert to previous versions, facilitating robust software engineering?
- What measures are in place to ensure the security of the code throughout the development process?
- How do debugging tools help identify coding errors, and what level of automated support is provided?
- Are there manual processes in place to double-check the automated input?
- How do automated testing frameworks ensure code functionality and catch bugs?
- Are there any manual processes in place to double-check the automated input for validation?
- How do simulation environments offer safe spaces to train and test AI system models, and what measures are in place to ensure the security and integrity of the simulation data?

Performance Analysis Tools:

- How do performance monitoring tools ensure optimal efficiency?
- What are the common KPIs measured?
- How do profiling tools analyze code performance and pinpoint efficiency bottlenecks?
- What KPIs are used to monitor performance?
- How can tools be used to predict potential software/hardware failures and perform maintenance proactively?

Machine Learning and AI Development:

- How are machine learning frameworks designed, and how are they used for developing intelligent systems?
- What measures are in place to ensure the quality of the algorithms and models produced?
- How do XAI tools assist developers in understanding how models make decisions?
- What measures are in place to ensure the transparency of the XAI tools used?

Security and Configuration Management:

- What are the measures and mechanisms put in place to secure the development and deployment tools?
- Is there a documented security policy for these tools, and are there security controls in place?
- Is the configuration of these tools documented?
- Does the organization have change control processes for the configuration of these tools?

Integration and Functionality:

- Can the development and deployment tools be integrated with other systems in the organization?
- Has the integration been tested and documented?
- What functionality do the development and deployment tools provide?
- Are the tools designed specifically for developing and deploying AI systems? Do they provide all the necessary features to support AI systems?
- What measures are in place to ensure that the development and deployment tools support the automation of model deployment and accuracy?

Management and Documentation:

- How are the tools and libraries used throughout the AI development lifecycle audited and monitored to ensure they meet the project goals and satisfy industry standards?
- How is the dependency on third-party suppliers managed?
- What measures are put in place to ensure that any potential risks associated with third-party suppliers are identified, mitigated, and communicated?
- How were the vendors for the development and deployment tools chosen?

- Were selection criteria based on the quality, functionality, and performance of the tool that best fit the organizational needs?
- What are the costs of the development and deployment tools?
- Are the costs within the organization's budget?
- Are there ways to optimize the spending without compromising the quality and reliability of the AI system?
- What training and support do the vendors provide?
- Is there adequate documentation and training to ensure that users can effectively use the tools?
- Is there adequate documentation of the development and deployment tools used for the AI system?
- Is the documentation current and sufficient to understand the functionality and logic of the tools?

Appendix 23: Audit Questions "End-User Training and Documentation"

Training Materials:

- Is the purpose of the intelligent system accurately described in the end-user documentation and training manuals, and is it suitable for its intended purpose?
- Are the user documentation and training materials comprehensive and well-designed, covering all critical aspects of the intelligent system, including data protection, privacy, information security, and ethical considerations?
- Is the user documentation provided in an easy-to-understand manner?
- Do the training materials adhere to any regulatory requirements or industry standards?
- Are the training materials complete?
- Are the ethical principles of compliance, transparency, accountability, and fairness addressed in the training materials for the intelligent system?
- Does the training plan for use of the intelligent system include scenarios that account for any biases or challenges that may arise during use?
- Are the training objectives for the end-users of the intelligent system clearly defined, and do the training materials correspond to these objectives?
- Are the end-users provided with concise and understandable information about the intelligent system, including its capabilities, limitations, and risks?

- Are the training materials available in multiple formats that accommodate different learning styles and disabilities?
- Does the end-user training cover any relevant regulatory and compliance requirements, such as data protection, privacy, and ethical considerations?
- Do the training materials cover cybersecurity, such as identifying cybersecurity risks and policies to reduce these risks?
- Are the risks that may arise from new tech components, algorithms, or other updates to the intelligent system covered in end-user training?
- Are the end-users regularly assessed for their understanding of the intelligent system through tests, surveys, or feedback mechanisms, and are their responses recorded and evaluated?
- Are incidents connected to improper end-user training properly addressed by updating training materials?

KPIs and Metering:

- Are user feedback metrics recorded and evaluated to assess how effectively users understand and use the intelligent system?
- Is training effectiveness measured and regularly evaluated?
- Are the methods used to deliver training and document progress maintained in accordance with organizational policies and regulatory guidelines?
- What metrics are used to evaluate the effectiveness of the end-user training and documentation? Are these metrics clearly defined, and how are they measured?

Appendix 24: Audit Questions for "Deployment and Monitoring"

Safety and Security:

- What incident response strategies are in place for managing AI system model failures and security breaches?
- Can preparedness and efficacy of both technical and organizational response strategies be demonstrated, including response times and communication methods?

- Are there established procedures for mandatory disclosure (in case of severe incidents) as required by laws, and how are they implemented?
- What measures has the organization implemented to manage access control and ensure that only authorized individuals can alter the intelligent system?
- How is proactive monitoring for intrusion detection performed, and what technologies or methodologies is the organization using?
- How are alerts managed and escalated? Who is responsible for escalation?
- What is the frequency and depth of vulnerability scans, and how are patches prioritized and applied?
- What contingency plans are in place for unexpected critical failures or extreme scenarios?
- What are the backup plans to handle catastrophic failures and ensure continuous service under adverse conditions?
- How is a continuous evaluation conducted to monitor real-time performance and detect drift in the objectives of the AI system model?
- How are drifts detected? How are they communicated and evaluated, and who is responsible?

Documentation and Control:

- How does release management in regulated environments comply with external and internal standards?
- How is compliance with new versions of an intelligent system ensured? How is it tested?
- What version control practices are in place to ensure transparency and accountability in updates and modifications?
- Are the extent and detail of documentation of change logs satisfactory for audit and compliance demands?
- How comprehensive and current is the documentation related to the intelligent system's deployment and operational procedures?
- Is all documentation up to date, accessible, and auditable?
- What are possible restrictions or requirements for access to corresponding documentation?

Appendix 25: Audit Questions "Decommissioning"

System Identification:

- What is the specific intelligent system being decommissioned?
- What are the primary functions and use cases of this intelligent system?
- How will decommissioning the intelligent system affect current business processes and operations?
- Are there any critical dependencies on this intelligent system that need to be addressed?
- Have all relevant stakeholders been informed about the decommissioning process?
- How are stakeholder concerns and feedback being managed?

Rationale for Decommissioning:

- Why is the intelligent system being decommissioned?
- Have alternative solutions or systems been identified and evaluated?

Governance for Decommissioning:

- Is there a governance committee in place to authorize decommissioning? Do the meeting minutes/records reflect who is authorized and empowered to make the decision?
- Who is responsible for the ultimate decision to decommission (expected or unexpected) an intelligent system? Did senior management sign off?
- How does the organization ensure continuity in decision-making processes that depend on the intelligent system?
- What alternative processes or systems are in place to maintain business operations seamlessly?

Data Management and Security:

- What procedures are in place for data migration during decommissioning?
- How is data securely disposed of or de-identified once it is no longer needed?
- How does the organization handle access control and user privileges during the decommissioning phase?
- How is unauthorized access during decommissioning prevented?
- Is all critical data securely transferred or archived without loss or corruption?

- What measures are in place to verify the integrity and completeness of data during migration and disposal?
- What are the data retention policies, and how do they comply with relevant regulations?
- Are logs and metadata archived securely (for compliance and forensic analysis), and how long are they retained?

Emergency Planning and Resources:

- Is there a documented emergency decommissioning plan, and how often is it tested?
- What resources are available for unexpected decommissioning scenarios?
- How are the effectiveness of emergency plans evaluated and tested?

Training, Monitoring, and Compliance:

- What training and awareness programs are provided to staff involved in the decommissioning process?
- How does the organization monitor and document the decommissioning process for compliance and audit purposes?

Aftermath of a Decommissioned System:

- Verify the solution is marked as "inactive/decommissioned" in the intelligent system inventory of record, and any associated risk decisions in the risk register are confirmed to be "closed/not applicable" or "actionable," after decommissioning.
- Are any risks/issues closed that are no longer relevant?

Abbreviations and Glossary

AC	Alternating Current
AI	Artificial Intelligence
AI RMF	Artificial Intelligence Risk Management Framework
AIBOM	AI Bill of Materials
API	Application Programming Interface
APT	Advanced Persistent Threats
AR	Augmented Reality
ASCON	Family of lightweight authenticated ciphers
BCI	Brain Computer Interface
BLE	Bluetooth Low Energy
BOM	Bill of Materials
CAIO	Chief Artificial Intelligence Officer
CAO	Chief Audit Officer
CCPA	California Consumer Privacy Act
CIA	Confidentiality, Integrity, and Availability
CoAP	Constrained Application Protocol
CPRA	California Privacy Rights Act
CPU	Central Processing Unit
CSA	Cloud Security Alliance
DC	Direct Current
DDS	Data Distribution Service

DMCA	Digital Millennium Copyright Act
DNA	Deoxyribonucleic Acid
DORA	Digital Operational Resilience Act
EMP	Electromagnetic Pulse
ENISA	European Union Agency for Cybersecurity
EU	European Union
EU AI Act	European Union Artificial Intelligence Act
FADP	Swiss Federal Act on Data Protection
GANs	Generative Adversarial Networks
GDPR	General Data Protection Regulation
GenAI RMF	Generative AI Risk Management Framework
GPU	Graphics Processing Unit
HDD	Hard Disk Drive
HSM	Hardware Security Modules
IDE	Integrated Development Environment
IDS	Intrusion Detection Systems
IoT	Internet of Things
ISO	International Organization for Standardization
KPI	Key Performance Indicator
KRI	Key Risk Indicator
LLM	Large Language Model
LoRaWAN	Long Range Wide Area Network
MFA	Multi-factor Authentication

NB-IoT	Narrowband Internet of Things
NIS2	Network and Information Security Directive 2
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
OPC UA	OPC Unified Architecture
OSHA	Occupational Safety and Health Administration
PCA	Principal Component Analysis
PII	Personally Identifiable Information
RAM	Random Access Memory
RFID	Radio Frequency Identification
SAIBOM	Software and AI Bill of Materials
SBOM	Software Bill of Materials
SEC	Securities and Exchange Commission
SEO	Search Engine Optimization
SIEM	Security Information and Event Management
Sigfox	Low-power, long-range wireless communication protocol
SMPC	Secure Multi-Party Computation
SSD	Solid State Drives
TEE	Trusted Execution Environments
TOS	Terms of Service
TPU	Tensor Processing Unit
UI	User Interface
VDP	Vulnerability Disclosure Program

VR	Virtual Reality
WLAN	Wireless Local Area Network
XAI	Explainable AI
Zigbee	Low-power, short-range wireless communication protocol
Z-Wave	Wireless protocol for home automation

CSA provides an Online Glossary [49].

Bibliography

[1]	World Digital Technology Academy (WDTA), "Generative AI Application Security Testing and Validation Standard," World Digital Technology Academy (WDTA), 04 2024. [Online]. Available: https://wdtacademy.org/publications/GenerativeAiApplicationSecurityTestingAndValidationStandard. [Accessed 14 08 2024].
[2]	L. Floridi, M. Holweg, M. Taddeo, J. Amaya, J. Mökander and Y. Wen, "capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act," Center for Digital Ethics.
[3]	NIST, "NIST AI RMF Playbook," National Institute of Standards and Technology (NIST), [Online]. Available: https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook. [Accessed 14 08 2024].
[4]	OWASP Top 10 for LLM Applications Team, "LLM AI Cybersecurity & Governance Checklist," OWASP, 2024.
[5]	World Digital Technology Academy (WDTA), "Large Language Model Security Testing Method," World Digital Technology Academy (WDTA), 04 2024. [Online]. Available: https://wdtacademy.org/publications/LargeLanguageModelSecurityTestingMethod. [Accessed 14 08 2024].
[6]	J. Mökander, J. Schuett, H. R. Kirk and L. Floridi, "Auditing large language models: a three-layered approach," *AI and Ethics,* 05 2023.
[7]	D. Stocker, J. Martella, A. Sharpe and I. Okoli, "Don't Panic! Getting Real About AI Governance," Cloud Security Alliance (CSA), [Online]. Available: https://cloudsecurityalliance.org/artifacts/dont-panic-getting-real-about-ai-governance. [Accessed 18 09 2024].
[8]	ISO, "ISO/IEC 42001:2023," 2023. [Online]. Available: https://www.iso.org/standard/81230.html. [Accessed 14 08 2024].
[9]	ISO, "ISO/IEC 5338:2023," 2023. [Online]. Available: https://www.iso.org/standard/81118.html. [Accessed 14 08 2024].
[10]	C. Spleiss, R. Ayalin, F. Chyla, B. Gaylord, F. Haenig, R. Heckman, H. Labib, L. Ruddikeit, A. Sharpe and A. Vashishtha, "AI Resilience: A Revolutionary Benchmarking Model for AI Safety," Cloud Security Alliance (CSA), 2024.

[11]	C. Spleiss, "AI Resilience & Diversity - BLOG," Cloud Security Alliance (CSA), 2024.
[12]	J. Mökander, "Auditing of AI: Legal, Ethical and Technical Approaches," *Digital Society (DISO),* vol. 2 , no. 49 (2023), 08 11 2023.
[13]	ISO, "Artificial intelligence concepts and terminology," [Online]. Available: https://www.iso.org/standard/74296.html. [Accessed 02 10 2024].
[14]	P. Phillips, C. Hahn, P. Fontana, A. Yates, K. Greene and D. a. P. M. Broniatowski, "Four Principles of Explainable Artificial Intelligence," 2021.
[15]	imda & pdpc - Singapore, "Model Artificial Intelligence Governance Framework," 21 01 2020. [Online]. Available: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf. [Accessed 13 08 2024].
[16]	CSA Workgroup AI Governance and Compliance, "AI Governance & Compliance Resource Links Hub," Cloud Security Alliance (CSA), [Online]. Available: https://cloudsecurityalliance.org/ai-governance-compliance-resource-links. [Accessed 13 08 2024].
[17]	ISO, "ISO/IEC 27001:2022," [Online]. Available: https://www.iso.org/standard/27001. [Accessed 13 08 2024].
[18]	NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," National Institute of Standards and Technology, 01 2023. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. [Accessed 13 08 2014].
[19]	NIST, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile: NIST AI 600-1," National Institute of Standards and Technology, 07 2024. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf. [Accessed 13 08 2024].
[20]	Cyber Risk GmbH, "The Framework for AI Cybersecurity Practices (FAICP)," [Online]. Available: https://www.faicp-framework.com/. [Accessed 13 08 2024].
[21]	United Nations, "THE 17 GOALS," [Online]. Available: https://sdgs.un.org/goals. [Accessed 13 08 2024].
[22]	NASA, "The Unknown Known," *System Failures Case Studies,* vol. 3, no. 01, 01 2009.
[23]	Digital Curation Centre Trilateral Research, "The Role of Data in AI," 2020.

[24]	"General Data Protection Regulation (GDPR)," EUR-Lex, 27 04 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj.
[25]	GDPR.EU, "Data Protection Impact Assessment (DPIA)," Proton Technologies AG, [Online]. Available: https://gdpr.eu/data-protection-impact-assessment-template/. [Accessed 13 08 2024].
[26]	CA.GOV, "California Privacy Protection Agency," [Online]. Available: https://cppa.ca.gov/. [Accessed 13 08 2024].
[27]	Federal Council of Switzerland, "Federal Act on Data Protection (FADP)," [Online]. Available: https://www.fedlex.admin.ch/eli/cc/2022/491/en. [Accessed 13 08 2024].
[28]	U.S. Copyright Office, "The Digital Millennium Copyright Act," [Online]. Available: https://www.copyright.gov/dmca/. [Accessed 13 08 2024].
[29]	R. Matulionyte, "Researchers warn we could run out of data to train AI by 2026. What then?," 07 11 2023. [Online]. Available: https://theconversation.com/researchers-warn-we-could-run-out-of-data-to-train-ai-by-2026-what-then-216741. [Accessed 13 08 2024].
[30]	P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu and L. H. M. Heim, "Will we run out of data? Limits of LLM scaling based on human-generated data," 04 06 2024. [Online]. Available: https://arxiv.org/abs/2211.04325. [Accessed 13 08 2024].
[31]	R. Elgart, "The Data Black Market: Where Hackers Take Stolen Data," 08 05 2019. [Online]. Available: https://www.turn-keytechnologies.com/blog/article/the-data-black-market-where-hackers-take-stolen-data. [Accessed 13 08 2024].
[32]	Y.-w. Kim, "How Transferable are Video Representations Based on Synthetic Data?," MIT, 2022.
[33]	X. Lin, J. Liu, J. Hao, K. Wang, Y. L. H. Zhang and H. a. T. X. Horimai, "Collinear holographic data storage," *Opto-Electronic Advances,* vol. 3, no. 3, 2020.
[34]	L. Ceze and J. &. S. K. Nivala, "Molecular digital data storage using DNA," *Nature Reviews Genetics,* vol. 20, p. 456–466, 2019.
[35]	R. Grimes, E. Chiu, J. Gable, B. Huttner and L. Perret, "Practical Preparations for the Post-Quantum World," Cloud Security Alliance (CSA), 2021.
[36]	NIST, "Lightweight Cryptography," National Institute of Standards and Technology, [Online]. Available: https://csrc.nist.gov/Projects/lightweight-cryptography. [Accessed 14 08 2024].

[37]	T. Liu and G. a. J. R. Ramachandran, "Post-Quantum Cryptography for Internet of Things: A Survey on Performance and Optimization," 2024.
[38]	European Commission, "EU AI Act," European Commission, 2021.
[39]	Council of the EU , "Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world," 09 12 2023. [Online]. Available: https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/. [Accessed 14 08 2024].
[40]	Colorado General Assembly, "Consumer Protections for Artificial Intelligence: SB24-205," 2024. [Online]. Available: https://leg.colorado.gov/bills/sb24-205. [Accessed 14 08 2024].
[41]	J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost and N. a. L. S. Wiebe, "Quantum Machine Learning," 2018.
[42]	C. Schuman, S. Kulkarni, M. Parsa, P. J. Mitchell, P. Date and B. Kay, "Opportunities for neuromorphic computing algorithms and applications," *Nature Computational Science* , vol. 2, p. 10–19, 31 01 2022.
[43]	N. Barney, "What is neuromorphic computing?," TechTarget, 04 2023. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/neuromorphic-computing. [Accessed 14 08 2024].
[44]	U.S. Department of Labor, "Hazards Associated with Industrial Robot Applications," U.S. Department of Labor, [Online]. Available: https://www.osha.gov/otm/section-4-safety-hazards/chapter-4#hazards. [Accessed 13 08 2024].
[45]	Harvard University Privacy Tools Project, "Differential Privacy," [Online]. Available: https://privacytools.seas.harvard.edu/differential-privacy. [Accessed 14 08 2024].
[46]	K. Martineau, "What is federated learning?," IBM, 24 08 2022. [Online]. Available: https://research.ibm.com/blog/what-is-federated-learning. [Accessed 14 08 2024].
[47]	Confidential Computing Consortium, "About the Confidential Computing Consortium," [Online]. Available: https://confidentialcomputing.io/about/. [Accessed 14 08 2024].
[48]	S. Burke, M. Capotondi, D. Catteddu and K. Huang, "Large Language Model (LLM) Threats Taxonomy," Cloud Security Alliance (CSA), 2024.

[49]	CSA, "Cloud Security Glossary," Cloud Security Alliance (CSA), [Online]. Available: https://cloudsecurityalliance.org/cloud-security-glossary. [Accessed 14 08 2024].