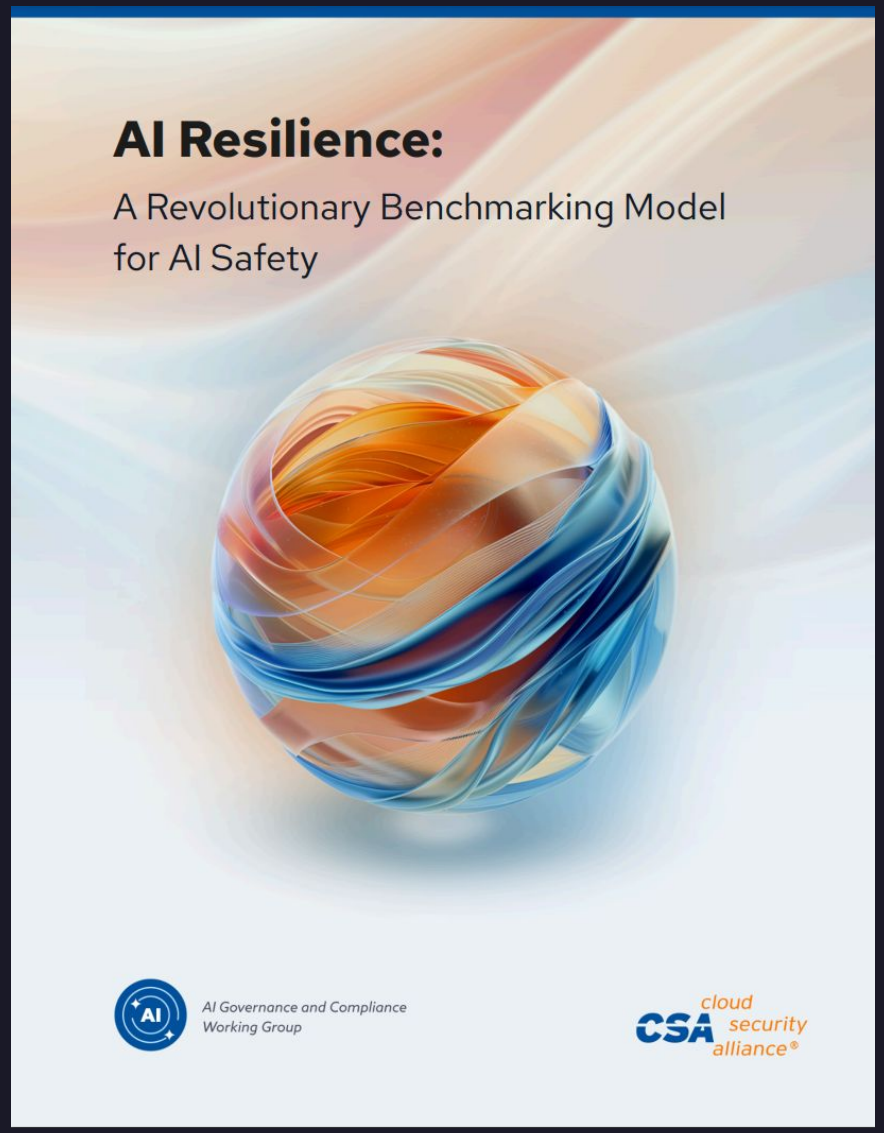
AI Resilience: A Revolutionary Benchmarking Model for AI Safety

Release Date: May 6th, 2024

cloud security alliance®

AI Governance and Compliance Working Group



Acknowledgements

Lead Authors

Dr. Chantal Spleiss

Contributors

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

CSA Global Staff

Ryan Gifford

Reviewers

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

Executive Summary

- Introduction of Benchmarking Model
  - New AI benchmarking model to navigate AI governance and compliance.
  - Inspired by evolution and psychology principles.
  - Prioritizes robustness alongside performance.

- Industry Analysis and Case Studies
  - Insights from automotive, aviation, critical infrastructure, defense, education, finance, and healthcare.
  - Lessons from past AI failures.

- Advocacy for Ethical AI
  - Integrates diverse perspectives with regulatory guidelines.
  - Focuses on trustworthiness to minimize risks and protect reputation.

- Empowerment of Decision Makers
  - Aids government officials, regulatory bodies, and industry leaders.
  - Establishes frameworks for ethical AI development, deployment, and use.
  - Provides a tool for assessing AI quality and ensuring long-term success.

Introduction

- Challenges in AI Evolution
  - Rapid evolution poses escalating risks.
  - Incidents from biased algorithms to malfunctioning vehicles highlight AI failures.
  - Current frameworks struggle with technological innovation pace.

- Holistic Perspective on AI Governance
  - Urgent need for more comprehensive AI governance and compliance.
  - Novel approach compares AI evolution with biology.
  - Introduces diversity concept to enhance AI safety.

- Innovative Benchmarking Framework
  - Increases safety and reliability of AI technology.
  - Empowers decision-makers and technical teams to assess AI systems.

- Advocacy for Ethical AI
  - Integrates diverse perspectives and regulatory guidelines.
  - Aims to establish strong governance practices and foster ethical AI innovation.

cloud security alliance®

PART 1: Understanding the Foundations

AI Governance and Compliance Working Group

Governance vs. Compliance

- Essential Aspects of Management
  - Ensures adherence to regulations, ethical principles, standards, and sustainability practices.
  - Alignment with principles and regulations ensures business continuity and ethical practice.

- Governance Framework
  - Implemented in a top-down approach by senior management.
  - Defines strategy, risk appetite, and establishes policies, standards, and procedures.
  - Shapes risk management, compliance obligations, and decision-making.
  - Creates culture of accountability, transparency, ethical behavior, and sustainability.
  - Prioritizes security and privacy measures.

- Compliance Mechanism
  - Follows a bottom-up approach.
  - Employees adhere to governance framework to meet regulatory requirements.
  - Focuses on adherence to laws, regulations, and internal business code of conduct.
  - Ensures operation within legal, ethical boundaries, and minimized risk exposure.

Governance and Compliance: A Moving Target

Current Framework is based on these general requirements:

- **Human Oversight**
  - Control: Subject to human oversight and control.
  - Intervention: Mechanisms for human intervention when necessary.
  - Monitoring: Coupled with automated monitoring, scalable and practically applicable.

- **Safety and Reliability**
  - Prioritization: Minimize the risk of harm.
  - Methods: Through rigorous testing, validation, and risk assessment processes.
  - Mechanisms: Implementation of kill-switch or recourse mechanisms.

- **Ethical Considerations**
  - Adherence: To ethical principles, respect for human rights, and promotion of fairness.

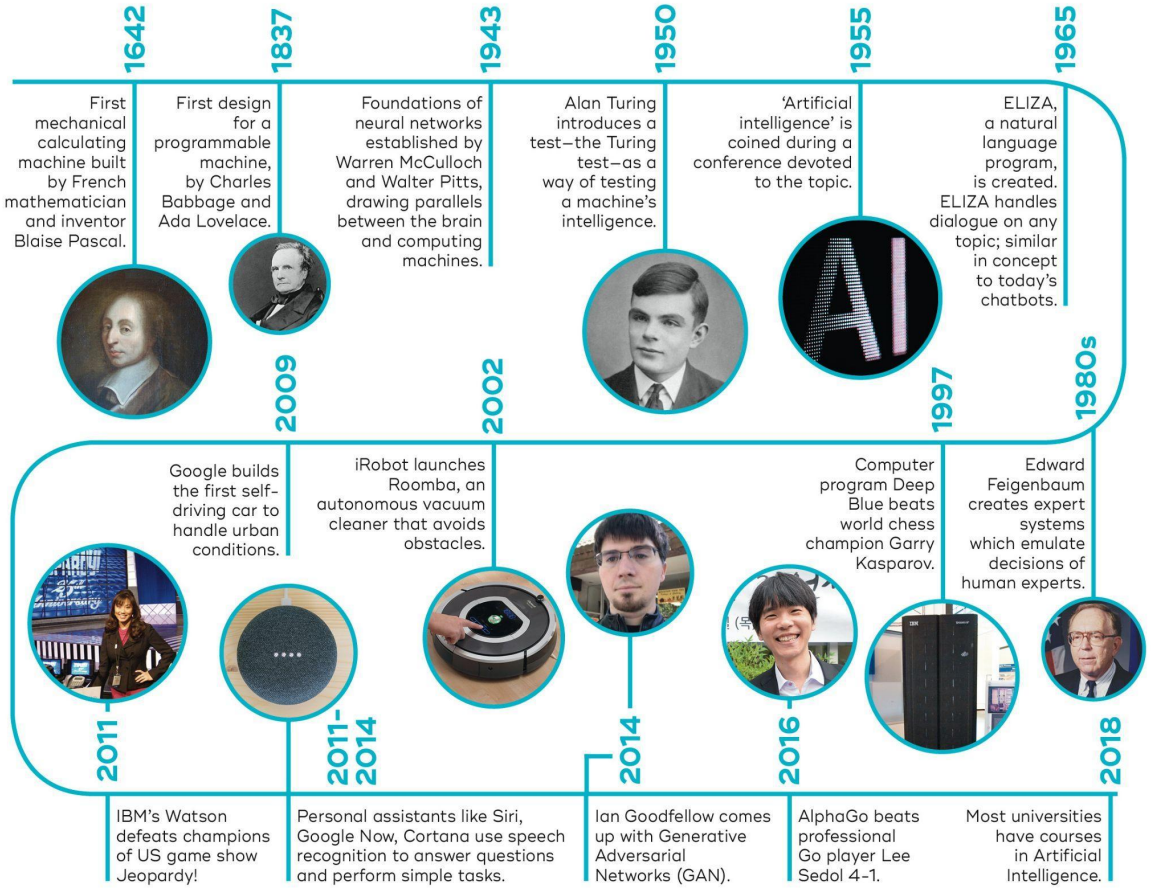
- **Data Privacy and Security**
  - Protection: Enhanced data protection and security measures.
  - Focus: Privacy-by-design and security-by-design to mitigate risks early.

- **AI Model and Data Considerations**
  - Bias Mitigation: Regularly monitor and evaluate for bias and discrimination.
  - Transparency: Explain workings, including algorithms; implementing XAI to foster trust.
  - Consistency: Ensures learning from accurate and reliable examples.
  - Accountability: Clear lines of responsibility; shared responsibility highlighted through manuals or "model cards".
  - Robustness: Resilient to adversarial attacks and other interferences.
  - Compliance: With laws, regulations, and standards including data protection, privacy, and safety.

The Landscape of AI

A Brief History of AI:

- 2018: BERT: Introduced by Google, this model revolutionized language understanding. BERT's use of the Transformer architecture and pre-training on massive text data sets enabled it to outperform previous models in various language tasks.
- 2019: GTP-2 with 1.5 billion parameters.
- 2020: LLMs with 175 billion - 530 billion parameters.
- 2021: LLMs with up to a trillion parameters focusing on improving efficiency in training and handling complex tasks with advanced reasoning and factual accuracy.
- 2022: ChatGTP-3 goes viral.
- Beyond size: researchers are working now on efficiency in training, alignment with human's value, safety and multimodality (incorporating images, audio, and other data types).
- This brief history of AI demonstrates the evolution from the most basic calculator to GenAI with Artificial General Intelligence still on the horizon.



The Landscape of AI

Different AI technologies are presented and discussed:

- **Machine Learning (ML)**
  - Machine Learning is a branch of AI and computer science that focuses on using data and algorithms to imitate human learning, gradually improving a model's accuracy.

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

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

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

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

The Landscape of Training Methods

- Supervised Learning
  - Learning from "labeled data" for classification and regression problems.
  - Examples:
    - Classification: Decision trees, random forests, linear classifiers, support vector machines.
    - Regression: Linear regression, multivariate regression, regression trees, lasso regression.

- Unsupervised Learning
  - Algorithms analyze unlabeled data to discover patterns or insights without predetermined outcomes.
  - Examples: k-means, k-medoids, hierarchical clustering, Apriori, FP Growth.

- Reinforced Learning
  - Learning through trial and error as an agent interacts with an environment, receiving rewards or penalties.
  - Examples: Markov Decision Process, Q-learning, Policy Gradient Method, Actor-Critic.

- Semi-supervised Learning
  - Uses a small amount of labeled data with a larger pool of unlabeled data.
  - Valuable when obtaining labeled data is costly or time-consuming.

- Self-supervised Learning
  - Generates its own labels from raw data through techniques like predicting masked words or video frames.
  - Allows learning robust, generalizable representations without human-provided labels.

The Landscape of Training Methods

- Federated Learning
  - Trains algorithms across decentralized devices, preserving privacy by not exchanging data.
  - Addresses privacy, security, data centralization concerns; utilizes "wisdom of the crowd".
  - Risk of malicious users disrupting model accuracy or privacy; uses techniques like differential privacy to enhance security.

- Regulations and Ethical Considerations
  - Impacted by GDPR, EU AI Act, OECD principles on AI.
  - No specific regulations for ML training but evolving rapidly with technology.
  - Significant for data monetization, business decision guidance; influences operational and strategic adjustments.
  - Meeting regulations introduces significant costs, particularly for businesses across multiple jurisdictions.
  - Navigating regulations adeptly offers competitive advantages like enhanced security and ethical transparency.

Licensing, Patenting & Copyright of AI Technology

- Open Source Initiative Licensing
  - Examples: Apache 2.0, MIT.
  - Note: Certain licensing might forbid commercial use of the resulting application.

- European Patent Office's (EPO) Guidelines for Examination
  - Revised Guidelines: Recent amendments mandate that applicants for AI or ML inventions further elucidate mathematical techniques and training input/data.
  - Purpose: To replicate the technical result of the invention over the entirety of the claim.

- Japan Agency for Cultural Affairs (ACA)
  - Draft Released: "Approach to AI and Copyright" for public comment on January 23, 2024.
  - Public Interaction: After considering nearly 25,000 comments, additional changes were made.
  - Anticipated Adoption: Will likely be adopted by the ACA in the next few weeks.

cloud security alliance®

PART 2: Real-World Case Studies and Industry Challenges

AI Governance and Compliance Working Group

A Brief History of AI Case Studies (2016–2019)

- 2016: Microsoft's Tay
  - AI chatbot transformed into a platform for racist remarks within 24 hours.
  - Microsoft deleted offensive content and shut down Tay.
  - Highlights the need for iterative improvement and proactive measures in AI design.

- 2018: Amazon's AI Recruiting Tool
  - Machine-learning engine favored male candidates.
  - Amazon disbanded the project due to fairness concerns.
  - AI in recruitment processes has become mainstream, albeit cautiously.

- 2019: Tesla Autopilot Accidents
  - Fatal collision involving Autopilot on a Tesla Model 3.
  - Misleading marketing and the system's limitations highlighted.
  - Raises questions about accountability and public safety in autonomous driving.

- 2019: Healthcare Algorithm Racial Bias
  - Algorithm inaccurately predicts health risks for Black patients compared to White patients.
  - Bias stems from using healthcare costs as a proxy for health needs.
  - Identifying more Black patients for additional care suggested as a remedy.

- 2019: Allegations of Apple Card Bias
  - Viral Twitter thread leads to regulatory investigation of Goldman Sachs' credit card practices.
  - Differences in credit lines for male and female customers noticed.
  - Investigation finds no discrimination based on sex.

A Brief History of AI Case Studies (2020-2024)

- 2020: Biased Offender Assessment Systems
  - Tools like COMPAS and OASys criticized for lack of transparency and fairness.
  - Assist in sentencing, probation, and treatment programs.
  - Criticism centers around algorithm transparency, fairness, and biases.

- 2022: Air Canada Chatbot's Misleading Info
  - Chatbot provided incorrect information about the airline's bereavement travel policy.
  - Air Canada argued chatbot operated independently.
  - Tribunal ruled in favor of the passenger, highlighting AI accountability.

- 2023: Lawsuit Against UnitedHealth's AI
  - AI allegedly denies coverage for elderly patients' essential care.
  - Overrides doctors' recommendations, sparking concerns about patient well-being.
  - Lawsuit underscores the need for transparency and human oversight in healthcare AI.

- 2024: Google's Gemini: A Lesson in AI Bias
  - Gemini 1.5 chatbot criticized for generating biased images.
  - Elon Musk and conservatives accuse Google of biased algorithms.
  - Incident prompts debates on AI ethics, diversity initiatives, and algorithmic accountability.

Industries: Regulations and Challenges

Automotive

- AI Implementation in Automotive:
  - Focus on automated and autonomous driving (SAE level 4 and 5).
  - Emphasis on safety for onboard systems and components.

- Current and Upcoming Standards:
  - Several ISO standards mention or partly regulate AI.
  - New standards drafted or under review by various regulatory bodies.

- Regulatory Impacts:
  - Regulation (EU) 2019/2144 mentions automated vehicles, with specific safety requirements for AI systems in vehicles.
  - ISO PAS 8800 focuses on drafting AI safety principles for road vehicles.

- AI Safety and Functional Standards:
  - ISO/TR 5469:2024 describes AI functional safety in automotive applications.
  - ISO/TR 4804:2020 emphasizes cybersecurity for automated driving systems, to be replaced by ISO/CD TS 5083.

Industries: Regulations and Challenges

Aviation

- AI Standards and Security in Aviation:
  - Aviation adopts IT security standards like ISO/IEC 27001, ISO/IEC 42001, ISO/TR 5469, NIST AI RMF, and AI ethics standards.
  - No AI-specific regulations enforced yet.

- Global Aviation Bodies and AI:
  - Governing bodies such as US FAA, EU EASA, UK CASA, and AU CASA recognize AI's potential and challenges but haven't regulated its use yet.
  - UK CASA in request-for-response stage with industry surveys.
  - US FAA has an AI technical discipline team led by Dr. Trung T. Pham.

- AI Applications in Military and Civil Aviation:
  - Military uses include intelligence analysis, autonomous vehicles, predictive maintenance, and security.
  - Civil uses focus on weather planning, route optimization, maintenance, and management of passengers and cargo.
  - Generative AI used in customer chatbots and decision support systems.

- Research and Future Directions:
  - Significant research on AI for air traffic control, including EU's CORDIS Results Pack on AI in air traffic management (October 2022).
  - Continuous updates in AI regulations needed due to technology advancements and the long lifespan of commercial airliners.

Industries: Regulations and Challenges

Critical Infrastructure & Essential Services

- Shift Towards AI-Driven Systems:
  - Significant shift towards more efficient, responsive, and intelligent systems.
  - Sectors impacted include electricity, gas, water, and food supply chains.

- Challenges and Opportunities:
  - Embracing digital transformation while balancing performance enhancement and security robustness.
  - Focus on the importance of regulatory frameworks, security standards, and the need for continuous adaptation.

- Performance vs. Security:
  - High-performing AI systems promise improved efficiency and operational optimization.
  - Introduction of vulnerabilities through IoT devices with integrated tinyML and edge computing.
  - Expansion of attack surface due to decentralizing data processing.

- Regulatory Frameworks and Standards:
  - Development of general frameworks like ISO/IEC 27001, ISO/IEC 27002, ISA/IEC 62443 series, IEC TS 62351-100-4:2023, and IEC TR 61850-90-4:2020.
  - Need for specific regulations targeting IoT and edge AI remains unclear.

Industries: Regulations and Challenges

Critical Infrastructure & Essential Services

- IoT and Edge AI Risks:
  - Integration of IoT devices introduces risks of cyberattacks.
  - Devices could be exploited to manipulate AI-driven decisions, impacting essential services.

- Future-Proofing Infrastructure:
  - Development of sector-specific AI regulations.
  - Standardized security protocols for IoT and edge AI crucial for protection against cyber threats.
  - International collaboration on AI governance to ensure effective global security measures.

- Continuous Evolution and Adaptation:
  - Navigating the balance between innovation and security through vigilant, adaptive governance.
  - Importance of ongoing training, sharing of best practices, and embracing "never trust, always verify" principles.
  - Strategic preparation against threats and leveraging AI in defense mechanisms.

- Strategic Initiatives and Global Regulations:
  - US Executive Order 14110 and EU AI Act outline frameworks for managing AI risks in critical infrastructure.
  - OECD AI Principles emphasize accountability in AI development and deployment.
  - Artificial Intelligence and Data Act (AIDA) guides the responsible use of AI within Canada, emphasizing safety, accountability, and the mitigation of harms. Outlines roles and obligations of stakeholders and enforcement mechanisms to ensure compliance.

Industries: Regulations and Challenges

Defense

- AI as a Strategic Enabler:
  - AI intertwines physical and digital domains in future battlefields.
  - Key for situational awareness, intelligence, and decision-making.

- Integrating Technologies:
  - Robotics, autonomous systems, data, and biotechnology create opportunities and risks.
  - AI essential for exploiting these technologies and countering adversaries.

- Collaboration and Innovation:
  - Military partnerships with private sector, academia, and allies are crucial.
  - Promotes innovation, adoption, and development of new leadership roles.

- Regulations and Ethical Considerations:
  - Importance of AI regulations and frameworks for ethical, safe, and trustworthy use.
  - Advances defense industry by fostering innovation, collaboration, and enhancing public trust.

- Historical Developments:
  - Early AI investments driven by defense, primarily U.S. DARPA.
  - Evolution from semi-automated warfare to smart weapons and decision support systems.

- AI's Role in Modern Warfare:
  - U.S. and China's significant investments in AI for defense.
  - Potential for an AI arms race and autonomous warfare challenges.

- Ethical and Safety Concerns:
  - Risks of machines making life-and-death decisions.
  - Need for balance between autonomous decision-making and human oversight.

- Regulatory Landscape:
  - Defense-specific AI regulations are scant; reliance on public domain resources.
  - NATO and U.S. DOD initiatives focus on responsible and ethical AI use in military applications

Industries: Regulations and Challenges

Education

- Opportunities for Enhanced Learning:
  - AI technologies like adaptive learning systems, AI tutors, and predictive analytics offer personalized education.
  - Potential to enhance learning outcomes and address educational inequalities.
- Privacy and Data Protection Concerns:
  - Extensive data collection and processing by AI systems raise privacy concerns.
  - Importance of governance frameworks to ensure compliance with legal and ethical standards.
- Equity and Access:
  - Crucial to ensure AI tools do not worsen disparities but empower learners.
  - Addressing AI algorithm bias to prevent discriminatory outcomes.
- Collaborative Development:
  - Need for collaborations among educators, technologists, ethicists, and policymakers.
  - Transparent and inclusive development processes for ethically aligned AI systems.
- Community Engagement and Digital Literacy:
  - Continuous dialogue with stakeholders including students, parents, and educators.
  - Investing in digital literacy and AI education to prepare participants for effective engagement.
- Proactive and Nuanced Approach:
  - Balancing innovation with ethical considerations to safeguard rights and well-being.
  - Fostering collaborations, ensuring transparency, and prioritizing equity and access.

Industries: Regulations and Challenges

Finance

- Historical Context:
  - The financial industry is heavily regulated, adjusting continuously to new technology trends and consumer needs.
  - AI regulations have existed for years but are not widely recognized, classified under business risk in risk management.

- Early Signs from LTCM:
  - In 1994, Long-Term Capital Management (LTCM) led by Nobel Prize-winning economist Myron Scholes and top traders, specialized in arbitrage financial modeling.
  - LTCM faced a major crisis in 1998 due to the Russian debt default, contradicting their AI-driven models.
  - The U.S. government intervened with a $3.625 billion bailout to prevent a potential global financial crisis, leading to LTCM's liquidation in 2000.

- Regulatory Responses:
  - Post-LTCM, the Basel Committee on Banking Supervision in 2005 updated guidelines, influencing the use of AI in rating systems.
  - These systems, often powered by Deep Learning, assess creditworthiness and are crucial in modern banking yet remain opaque to users.

- Advancements in AI Regulation:
  - Following the 2008 financial crisis, the US Federal Reserve issued the "Supervision and Regulation Letters" SR 11-7 in 2011.
  - This guidance mandates banks with assets over $10 billion to adhere to strict model risk management, acknowledging the role of flawed AI models in the mortgage crisis.

Industries: Regulations and Challenges

Finance

Guidance on Model Risk Management SR 11-7

- Definition and Importance
  - In the US banking industry, SR 11-7 is akin to new legislation; compliance is mandatory.
  - Defines a model as a system applying statistical, economic, financial, or mathematical theories to process data into estimates.
  - Model risk involves potential adverse consequences from incorrect or misused model outputs, necessitating active management.

- Famous Investigations and Incidents
  - 2019 Apple credit card discrimination case highlighted issues with model validation.
  - $440M loss by Knight Capital Group in 2012 due to a failed software update, exemplifying the critical need for robust implementation and usage.

- Application Beyond Financial Products
  - Models aid in compliance with the Bank Secrecy Act and the US PATRIOT Act, crucial for anti-terrorism efforts.
  - Non-compliance can lead to severe penalties, including financial fines and restrictions on expansion.

- International Standards
  - The European Central Bank's revised Guide to Internal Models (2024) aligns with SR 11-7's principles, incorporating climate risks and default risk measurements.
  - Although Asia lacks specific guidelines like SR 11-7, model risk management practices are spreading globally.

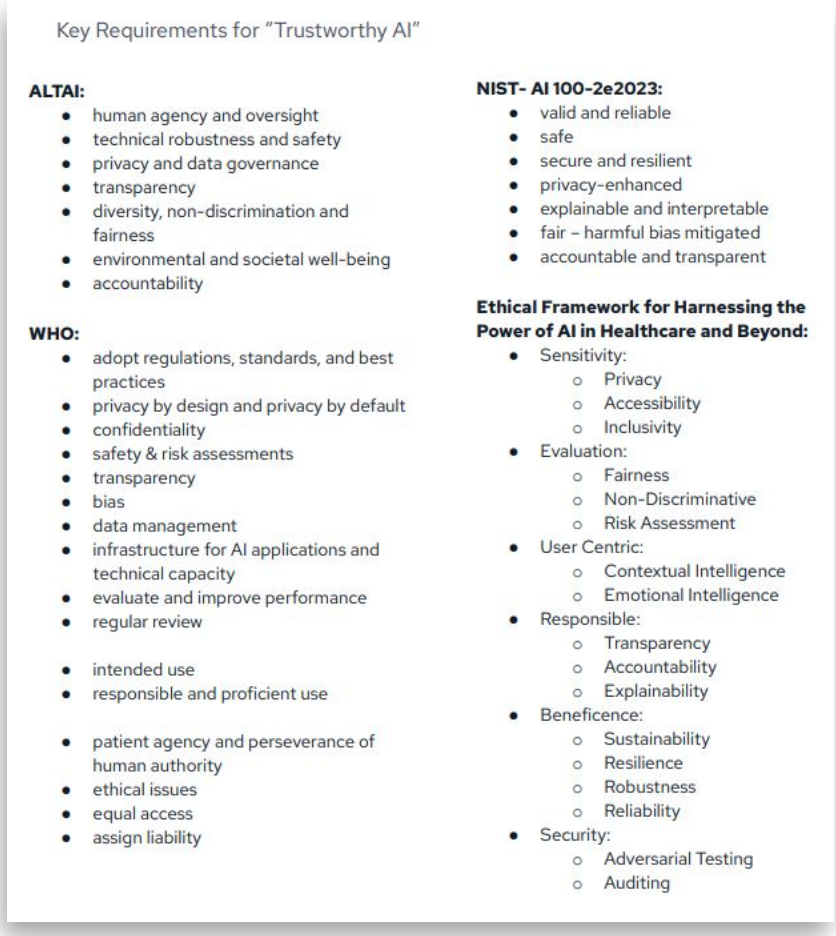
- Related Standards in Financial Industry
  - PCI DSS and PCI 3DS standards govern transaction data security and online transaction authentication, essential for preventing fraud.

- Innovation and Compliance
  - Financial industry innovation is driven by technology adoption; regulatory compliance remains a critical focus to protect data and customer privacy.

Industries: Regulations and Challenges

Healthcare

- AI Potential and Risks in Healthcare
  - Distinction between ML (Machine Learning) and GenAI (Generative AI) is crucial.
  - ML is more secure, being task-specific.
  - GenAI interacts with various stakeholders, raising issues in reliability, security, privacy, and misuse prevention.
- Regulatory Landscape and Trustworthy AI
  - Abundance of country-specific regulations, global standards, and best practices.
  - "Trustworthy AI" focuses on governance, compliance, and addressing technical challenges with a practical approach.
- Key Concepts of Trustworthy AI
  - AI to behave as intended and minimize risks.
  - Emphasis on explainability, reliability, security, privacy, and bias mitigation.
- Selected Literature on Trustworthy AI in Healthcare
  - Ethics and Governance of Artificial Intelligence for Health
  - Assessment List for Trustworthy Artificial Intelligence (ALTAI)
  - NIST Trustworthy and Responsible AI
  - Ethical Framework for Harnessing the Power of AI in Healthcare and Beyond
- Consolidated Requirements for Trustworthy AI
  - Includes human oversight, privacy, reliability, performance, transparency, diversity, fairness, technical robustness, and sustainability.
- Conclusions from Healthcare Literature
  - Various perspectives on AI requirements.
  - Highlight on the need for responsible AI use and liability considerations.
  - Surprising lack of emphasis on sustainability despite its relevance.


Key Requirements for "Trustworthy AI"

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

WHO:
- adopt regulations, standards, and best practices
- privacy by design and privacy by default
- confidentiality
- safety & risk assessments
- transparency
- bias
- data management
- infrastructure for AI applications and technical capacity
- evaluate and improve performance
- regular review
- intended use
- responsible and proficient use
- patient agency and perseverance of human authority
- ethical issues
- equal access
- assign liability

NIST- AI 100-2e2023:
- valid and reliable
- safe
- secure and resilient
- privacy-enhanced
- explainable and interpretable
- fair – harmful bias mitigated
- accountable and transparent

Ethical Framework for Harnessing the Power of AI in Healthcare and Beyond:
- Sensitivity:
  - Privacy
  - Accessibility
  - Inclusivity
- Evaluation:
  - Fairness
  - Non-Discriminative
  - Risk Assessment
- User Centric:
  - Contextual Intelligence
  - Emotional Intelligence
- Responsible:
  - Transparency
  - Accountability
  - Explainability
- Beneficence:
  - Sustainability
  - Resilience
  - Robustness
  - Reliability
- Security:
  - Adversarial Testing
  - Auditing

Industries: Regulations and Challenges

Healthcare

- Bias in Healthcare
  - Depersonalization of data is necessary to mitigate biases.
  - Balancing inclusion of characteristics and avoiding bias through Explainable AI (XAI).
  - Techniques like LIME and SHAP aid in post-hoc explainability.
  - XAI facilitates regulatory compliance and trustworthy AI in healthcare.

- Further Applications of ML/AI in Healthcare
  - Streamlining regulatory processes.
  - Optimizing supply chain management.
  - Assisting in drug development.
  - Improving direct and indirect patient care.
  - Enhancing in-home care with wearable devices and sensors.
  - Regulations and guidelines for developing medical devices with ML/AI.
  - Improving manufacturing processes in the pharmaceutical industry.

cloud security alliance®

PART 3: AI Resilience Reframed: Benchmarking Model Inspired by Evolution

AI Governance and Compliance Working Group

Comparison: Biological Evolution vs. AI Development

- Biological Evolution:
  - New features (mutations) tested for performance and resilience.
  - Organisms with persistence exhibit built-in protection against evolution.
  - Selection through different lenses enhances system capabilities.

- AI Development:
  - AI performance measured by output in a predefined context.
  - AI resilience includes generalization and adaptability.
  - Regulatory bodies tasked with overseeing safety and resilience.

- Future Trends:
  - Continuous learning AI systems set to dominate the market.
  - Dynamic systems require higher resilience compared to static ones.

- Challenges:
  - AI resilience often neglected in favor of performance.
  - Regulatory interventions become necessary for balancing innovation and regulation.

Diversity and Resilience in AI Systems

- Diversity in Problem-Solving:
  - Diversity mirrors nature's approach to problem-solving.
  - Encouraging and rewarding individual and unique AI approaches.
  - Diverse AI technologies with individual resilience enhance global security.

- Adaptability and Survival:
  - "It is not the strongest of the species that survives, nor the most intelligent, but the one most adaptable to change."
  - Applies to AI systems: resilience, not just performance, ensures survival.

- Enhancing AI Resilience:
  - Augmenting intrinsic resilience with user guardrails (e.g., manuals, training, warnings).
  - Technical prevention from "off-label" usage is crucial yet often overlooked.

- Prioritizing AI Resilience:
  - Policymakers, regulatory bodies, and governments must prioritize AI resilience.
  - Standardized metrics for evaluating resilience are essential for safe AI integrations.

AI Performance Benchmarking

- Approaching Saturation in Benchmarking:
  - AI benchmarking nearing saturation of traditional performance benchmarks.
  - Some systems surpassing human baseline performance.

- Stanford's Leadership in Benchmarking:
  - Stanford's Center for Research on Foundation Models leading with HELM.
  - Evaluates models across 87 scenarios and 50 metrics.

- Focus on Performance and Prevention of Harm:
  - Benchmarking focuses on performance and harm prevention.

- Resilience Evaluation:
  - Assessing resilience by testing model performance on diverse datasets (e.g., IMDB and BoolQ).
  - Emphasis on generalization while maintaining performance.

AI Resilience – Suggested Definition

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

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

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

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

The following definition of resilience for AI technology is suggested:

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

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

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

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

Proposed AI Resilience Score

- A resilience score from 0 to 10 is suggested, reflecting an AI's resilience based on its pillars: resistance, resilience, and plasticity.
- Example score: 16:5-8-3, indicating the sum of the three pillars and each separately.
- Distribution of scores reflects diversity among AI systems, aiding informed decision-making on risk mitigation.
- Policymakers, risk managers, regulatory bodies, and governments should prioritize AI resilience over performance, rewarding steps in this direction to promote diverse solutions and AI diversity.

Intelligence Awareness

- The concept of "Intelligence Awareness" emphasizes understanding differences in intelligence rather than comparing them.
- Intelligence Awareness is distinct from Howard Gardner's theory of multiple intelligences.
- It emphasizes the safe and efficient interaction between humans and intelligent systems by respecting diverse abilities.
- Illustrated by Andy Weir's "Project Hail Mary," it highlights the importance of adapting to different forms of intelligence.
- As AI nears or exceeds human capabilities, benchmarking becomes crucial, necessitating respect for diverse abilities.

Fundamental Differences in Intelligent Systems

- Fundamental Differences in Intelligence
  - Comparing AI and Human Intelligence (HI) assumes equivalency, with HI as the current standard.
  - Biological basis of HI vs. AI's silicon chip foundation significantly impacts functionality.

- Future Developments in AI
  - Potential for AI run on quantum or biological computers.
  - This integration could merge silicon chip efficiency with quantum capabilities akin to the human brain.

- Implications of Advanced AI Systems
  - AI could achieve performance of current systems plus human-like problem-solving skills.
  - Combination of deterministic and non-deterministic computing methods.

- Philosophical and Practical Considerations
  - How to evaluate an intelligence that produces incomprehensible solutions?
  - Reference to "42" from The Hitchhiker's Guide to the Galaxy highlights challenges in understanding AI outputs.

Additional Resources

- For more information about this topic, and to view a full comprehensive bibliography for the content provided in this presentation, please refer to the full "AI Resilience: A Revolutionary Benchmarking Model for AI Safety" document.