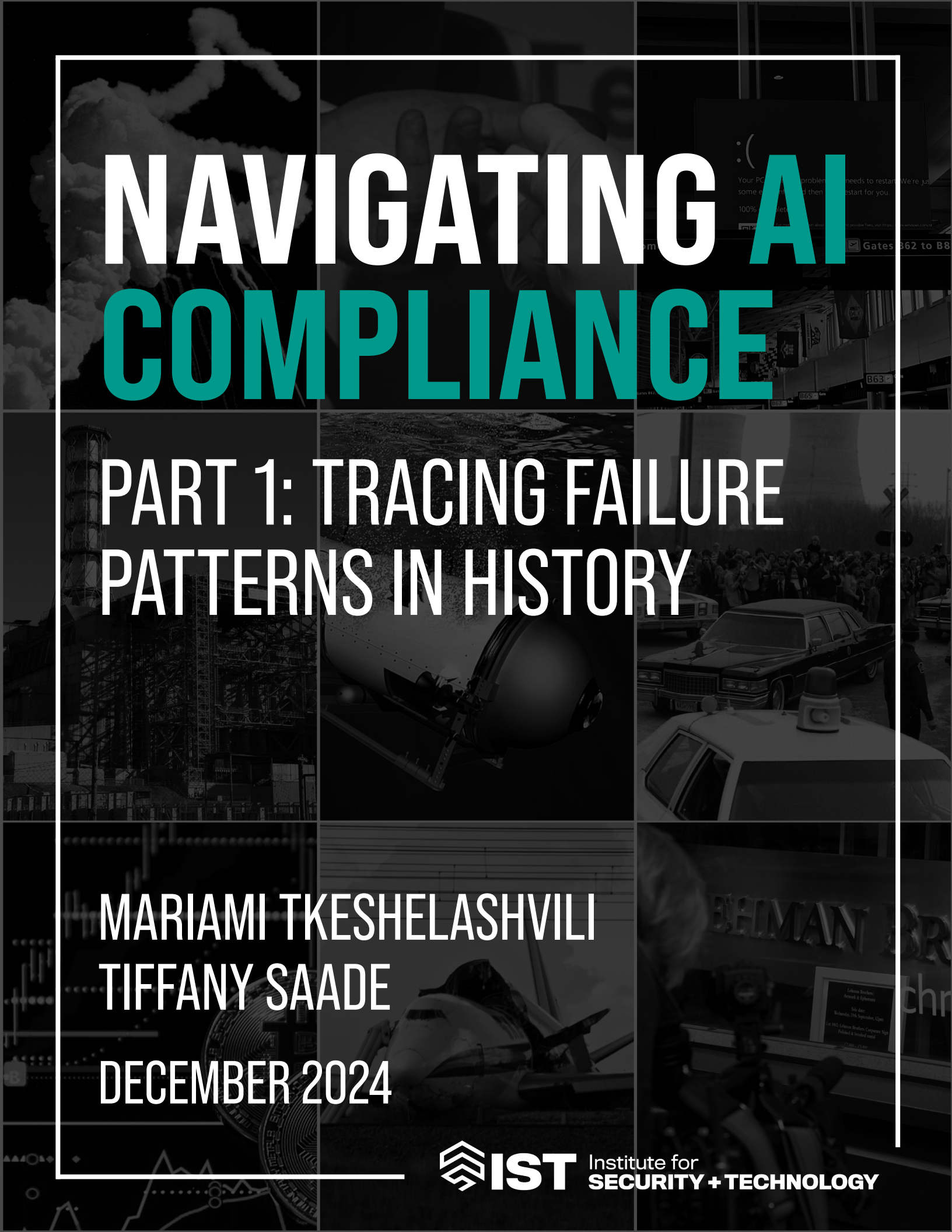


# NAVIGATING AI COMPLIANCE

## PART 1: TRACING FAILURE PATTERNS IN HISTORY

MARIAMI TKESHELASHVILI
TIFFANY SAADE

DECEMBER 2024

IST Institute for SECURITY + TECHNOLOGY

Cover page photo credit, from top left to bottom right:

The space shuttle Challenger explodes shortly after take-off, January 28, 1986. (NASA image #86-HC-220, Wikimedia Commons)

Theranos demonstrated getting blood drawn on demand at TechCrunch Disrupt SF, September 8, 2014. (Kevin Krejci, Flickr)

The Microsoft Windows Recovery screen displayed at Dulles Airport, July 19, 2024. (Reivax, Wikimedia Commons)

The damaged unit 4 reactor and shelter at Chernobyl, October 2010. (Dana Sacchetti/IAEA Imagebank)

3D model of the OceanGate submersible "Titan," June 2023 (Madelgarius, Wikimedia Commons)

President Carter leaves Three Mile Island nuclear power station after the accident, April 1, 1979. (Nuclear Regulatory Commission, Flickr)

Cryptocurrency illustration, January 4, 2021 (Illustration by Jorge Franganillo, CC)

A Boeing 747, operated by Kalitta Air, crashed in Brussels, May 25, 2008 (Simon Schoeters, Wikimedia Commons CC)

Lehman Brothers sign up for auction in London, UK, October 2010. (Jorge Royan, Wikimedia Commons)

# Navigating AI Compliance
# Part 1: Tracing Failure Patterns in History

December 2024

Authors: Mariami Tkeshelashvili, Tiffany Saade

Report Design: Sophia Mauro


Printed in the United States of America

IST

# About the Institute for Security and Technology

## *Uniting technology and policy leaders to create actionable solutions to emerging security challenges*

Technology has the potential to unlock greater knowledge, enhance our collective capabilities, and create new opportunities for growth and innovation. However, insecure, negligent, or exploitative technological advancements can threaten global security and stability. Anticipating these issues and guiding the development of trustworthy technology is essential to preserve what we all value.

The Institute for Security and Technology (IST), the 501(c)(3) critical action think tank, stands at the forefront of this imperative, uniting policymakers, technology experts, and industry leaders to identify and translate discourse into impact. We take collaborative action to advance national security and global stability through technology built on trust, guiding businesses and governments with hands-on expertise, in-depth analysis, and a global network.

We work across three analytical pillars: the Future of Digital Security, examining the systemic security risks of societal dependence on digital technologies; Geopolitics of Technology, anticipating the positive and negative security effects of emerging, disruptive technologies on the international balance of power, within states, and between governments and industries; and Innovation and Catastrophic Risk, providing deep technical and analytical expertise on technology-derived existential threats to society.

Learn more: https://securityandtechnology.org/

# Acknowledgments

This work is inherently collaborative. As researchers, conveners, and facilitators, the Institute for Security and Technology (IST) is immensely grateful to the members of the AI Risk Reduction multi stakeholder working group for their insights, dedication, willingness to engage in honest and healthy debate, and the time that each of them generously volunteered to this effort

We are also immensely grateful for the generous support of the Patrick J. McGovern Foundation, whose funding allowed us to continue this project through the lens of IST's Applied Trust & Safety program.

AI is too vast a set of tools, capabilities, and communities for any one organization to manage the risks and opportunities on its own. This effort reflects the cross-sectoral, public-private efforts needed more broadly across the ecosystem to ensure AI is beneficial for us all. We extend our gratitude to the following experts who contributed to this paper by providing their feedback, guidance, and participation in the multi stakeholder meetings:

- Chloe Autio
- Matthew da Mota
- Hadassah Drukarch
- Avijit Ghosh
- Elena Gurevich
- Katherine Johnson
- Brian Judge
- Mahesh Dinkar Nayak
- Alexander Reese
- Alyssa Lefaivre Škopac
- Peter Slattery
- Akash Wasil

Finally, the authors extend their gratitude to Steve Kelly and Philip Reiner for the support and strategic guidance they provided in the drafting and refining of this report and to editors Sophia Mauro and Jennifer Tang.

# Contents


**Executive Summary** ..... **1**

**Introduction** ..... **3**

**Methodology** ..... **4**

**Historical Perspective** ..... **5**

*Lehman Brothers Bankruptcy* ..... *5*

*Cambridge Analytica Scandal* ..... *5*

*Theranos Scandal* ..... *6*

*Boeing 737 MAX Crisis* ..... *6*

*FTX Collapse* ..... *7*

*CrowdStrike Outage* ..... *8*

Key Lessons From The Past ..... 8

**Sources of AI Governance** ..... **11**

Laws & Regulations ..... 11

*Table 1: State Bills Signed Into Laws in 2024 Targeting AI Developers and Users* ..... *13*

Guidance ..... 13

Norms ..... 14

Standards ..... 14

Organizational Policies ..... 14

**Defining Compliance Failure in the AI Context** ..... **15**

**Current State of Play** ..... **16**

Data Privacy & User Consent ..... 16

Algorithmic Bias ..... 17

**Outlook** ..... **19**

*Ambiguous AI safety definitions and the rapid pace of development challenges governance and, potentially, AI adoption across regulated industries.* ..... *19*

*Interpretability challenges will hinder development of compliance mechanisms.* ..... *19*

*AI Agents will blur the lines of liability in the automated world.* ..... *19*

**Conclusion** ..... **20**

**Appendix: Case Studies in AI-Adjacent Industries** ..... **21**

# Executive Summary

History often rhymes and echoes through the present and future. Through this lens, we examine past compliance failures across various industries–from nuclear power to financial services–to illuminate potential pitfalls in the AI ecosystem, offering definitions, frameworks, and lessons learned to help AI builders and users navigate today's complex compliance landscape.

Our analysis of eleven case studies from AI-adjacent industries reveals three distinct categories of failure: institutional, procedural, and performance. Institutional failures stem from a lack of executive commitment to create a culture of compliance, establish necessary policies, or empower success through the organizational structure, leading to foreseeable failures. Meanwhile, procedural failures are the result of a misalignment between an institution's established policies and its internal procedures and staff training required to adhere to those policies. Finally, performance failure results from an employee's failure to follow an established process, or an automated system's failure to perform as intended, leading to an undesirable result.

By studying failures across sectors, we uncover critical lessons about risk assessment, safety protocols, and oversight mechanisms that can guide AI innovators in this era of rapid development. One of the most prominent risks is the tendency to prioritize rapid innovation and market dominance over safety. The case studies demonstrated a crucial need for transparency, robust third-party verification and evaluation, and comprehensive data governance practices, among other safety measures. Additionally, by investigating ongoing litigation against companies that deploy AI systems, we highlight the importance of proactively implementing measures that ensure safe, secure, and responsible AI development. Recent court cases teach a crucial lesson: compliance with privacy, anti-discrimination, and transparency laws must be foundational, not an afterthought.

Though today's AI regulatory landscape remains fragmented, we identified five main sources of AI governance—laws and regulations, guidance, norms, standards, and organizational policies—to provide AI builders and users with a clear direction for the safe, secure, and responsible development of AI. In the absence of comprehensive, AI-focused federal legislation in the United States, we define compliance failure in the AI ecosystem as the failure to align with existing laws, government-issued guidance, globally accepted norms, standards, voluntary commitments, and organizational policies–whether publicly announced or confidential–that focus on responsible AI governance.

The report concludes by addressing AI's unique compliance issues stemming from its ongoing evolution and complexity. Ambiguous AI safety definitions and the rapid pace of development challenge efforts to govern it and potentially even its adoption across regulated industries, while problems with interpretability hinder the development of compliance mechanisms, and AI agents blur the lines of liability in the automated world. As organizations face risks ranging from minor infractions to catastrophic failures that could ripple across sectors, the stakes for effective oversight grow higher. Without proper safeguards, we risk eroding public trust in AI and creating industry practices that favor speed over safety—ultimately affecting innovation and society far beyond the AI sector itself. As history teaches us, highly complex systems are prone to a wide array of failures. We must look to the past to learn from these failures and to avoid similar mistakes as we build the ever more powerful AI systems of the future.

# Introduction

As highly advanced artificial intelligence (AI) systems become increasingly integrated into critical aspects of society—from healthcare and finance to transportation and national security—policymakers and broader society are paying closer attention to the potential risks associated with their development and deployment. The Institute for Security and Technology's (IST) December 2023 report on the risks and opportunities of cutting-edge AI foundation models identified six risk categories and assessed how varying levels of model openness influence each.[1] In follow on work, IST proposed the "lifecycle approach" to AI risk reduction, presented a deep dive on the risk of malicious use—one of the six risk categories, and suggested specific technology and policy mitigation strategies for that risk.

Building further upon that foundation, this report undertakes a deep dive on the risk of compliance failure, another of the six risk categories identified in the December 2023 report, defined therein as "the inability or unwillingness to adhere to established safety procedures, verification mechanisms, and legal requirements." What are these requirements, and procedures that must be complied with, and who is behind them? The answer is not straightforward, as global AI governance approaches are rapidly evolving and, to date, present an incomplete and often fragmented patchwork of requirements. Perhaps, we can draw lessons learned from other, more mature, contexts?

This report, the first in a two-part series, aims to:

- Provide historical context regarding compliance failures in adjacent industries;
- Define and contextualize compliance failure within the AI ecosystem;
- Analyze the current state of play of compliance failure trends in the AI ecosystem and the unique challenges posed by AI systems in maintaining compliance; and
- Discuss potential implications of compliance failure in the AI ecosystem.

The second installment, slated for publication in early 2025, will propose actionable risk reduction strategies and recommend broader interventions by a variety of players in the AI ecosystem.

Through these reports, and the convenings and conversations that accompany their development, we aim to contribute to the ongoing dialogue on AI governance and provide valuable insights for policymakers, industry leaders, and researchers working to ensure

1 Zoë Brammer, "How Does Access Impact Risk?" Institute for Security and Technology, December 2023, securityandtechnology.org/ai-foundation-model-access-initiative/how-does-access-impact-risk/.

the optimal development and deployment of AI technologies. Understanding both the historical and current patterns of compliance failure enables us to build a predictive model to better anticipate where the AI ecosystem might struggle to manage risk and the potential implications. While we are cautious about making explicit predictions for the future, this report underscores the need for effective AI governance and robust AI safety frameworks, as the rapid pace of technological advancement, coupled with the complex interplay of human choices, introduces significant uncertainty into any long-term forecast.

# Methodology

Our research relied on a series of historical case studies; analysis of available laws, guidance, norms, and standards on AI governance; investigation of databases that reflect the current state of the compliance issues within the AI ecosystem; and over 20 expert interviews with AI labs, tech industry stakeholders, machine learning engineers, AI governance and policy experts, compliance officers, lawyers, university-based AI research centers, AI ethicists, and independent researchers. Complementing this research, IST convened two multi-stakeholder, closed-door discussions to gather further insights.

First, we examined some of the most consequential cases of compliance failure in adjacent industries (e.g., nuclear energy, biotechnology, manufacturing, financial services, cybersecurity) to distill the lessons learned and identify trends and behavioral patterns. Historical cases illuminated several categories of compliance failures that became the points of analysis for the "current state of play."

Second, we analyzed the following datasets and databases that compile information on AI risks, AI incidents, AI-related litigation, and monetary costs of compliance failure: the Massachusetts Institute of Technology's AI Risk Repository Database, AI Incidents Database, AI Litigations Database, the European Union's General Data Protection Regulation (GDPR) Enforcement Tracker, Transatlantic Tech Policy Tracker, and Federal Trade Commission (FTC) cases on AI, among other sources.[2,3,4,5,6,7]

2 "The AI Risk Repository," Massachusetts Institute of Technology, accessed November 1, 2024, https://airisk.mit.edu/.

3 "Welcome to the Artificial Intelligence Incident Database," AI Incidence Database, accessed November 1, 2024, https://incidentdatabase.ai/.

4 "DAIL – the Database of AI Litigation," Ethical Tech Initiative, George Washington University, accessed November 1, 2024, https://blogs.gwu.edu/law-eti/ai-litigation-database/.

5 "GDPR Enforcement Tracker," CMS Law, accessed November 1, 2024, https://www.enforcementtracker.com/.

6 "Transatlantic Tech Policy Tracker," Center for European Policy Analysis, accessed November 12, 2024, https://cepa.org/issues/technology-and-innovation/transatlantic-tech-policy-tracker/.

7 "Artificial Intelligence Cases and Proceedings," U.S. Federal Trade Commission, accessed November 12, 2024, https://www.ftc.gov/industry/technology/artificial-intelligence.

Finally, based on historical case studies and our assessment of the current state of play, we employed a predictive mental model to offer an outlook on the specific challenges of the AI ecosystem in managing compliance risks.

# Historical Perspective

Case studies from AI-adjacent industries help us understand how different entities failed to comply with safety, security, transparency, and accountability standards. By analyzing the causes for these compliance failures, we drew the lessons learned, distilled some key categories and common factors of failure, and conceptualized compliance failure in the AI context. Our analysis draws from the below prominent case studies, augmented by additional examples summarized in the Appendix.

## LEHMAN BROTHERS BANKRUPTCY

Lehman Brothers Inc., once the fourth-largest investment bank in the United States, filed for bankruptcy on September 15, 2008, marking the largest bankruptcy filing in U.S. history. The collapse was a pivotal moment in the 2008 financial crisis.[8,9,10] It contributed to a widespread economic downturn, loss of jobs, and erosion of public trust in financial institutions. Key failures included:

- **Excessive leverage**: Lehman maintained a dangerously high leverage ratio, at times exceeding 30:1.
- **Fraudulent claims**: The bank used an accounting maneuver, called "Repo 105," to temporarily remove toxic assets from its balance sheet, presenting a falsely optimistic picture of its financial condition.
- **Inadequate risk management**: Lehman continued to invest heavily in the subprime mortgage market despite clear signs of market deterioration.
- **Regulatory oversight lapses**: The SEC failed to properly monitor Lehman's activities and enforce existing regulations.

## CAMBRIDGE ANALYTICA SCANDAL

Cambridge Analytica, a political consulting firm, improperly obtained the personal data of millions of Facebook users without their consent, using it for political advertising during the 2016 U.S. presidential election and other campaigns.[11,12] The scandal led to increased scrutiny of data privacy practices, bankrupted Cambridge Analytica,

8 "Corporate Governance Failures: The Lehman Brothers: Case Study," Faster Capital, last updated June 20, 2024, https://fastercapital.com/content/Corporate-Governance-Failures--The-Lehman-Brothers--Case-Study.html.

9 Stuart Gilson, Kristin Mugford, and Sarah L. Abbott, "The Rise and Fall of Lehman Brothers," Harvard Business School Case 217-041, January 2017 (revised January 2019), https://www.hbs.edu/faculty/Pages/item.aspx?num=52147.

10 Anton Valukas, "Lehman Brothers Volume 1," US Bankruptcy Court Southern District of NYC, 2010, https://web.stanford.edu/~jbulow/Lehmandocs/VOLUME%201.pdf.

11 Colin Earl, "Learning from a $150 Billion Compliance Failure," *Security Today*, November 18, 2018, https://securitytoday.com/articles/2018/11/27/learning-from-a-150-billion-compliance-failure.aspx.

12 Joseph Simons et al., "82 3107 United States of America before the Federal Trade Commission Commissioner," 2019, https://www.ftc.gov/system/files/documents/cases/182_3107_cambridge_analytica_administrative_complaint_7-24-19.pdf.

damaged public trust in social media platforms, and resulted in significant financial and reputational costs for Facebook.[13,14] It also sparked global discussions about data protection regulations and the ethical use of personal information in political campaigns. Key failures included:

- **Unauthorized data collection**: The firm collected personal data through a third-party app without proper user consent.
- **Misuse of personal information**: The data was then used for purposes beyond what had been previously agreed upon by the users.
- **Inadequate data protection measures**: Facebook failed to ensure that user data was protected and not misused by third-party applications.
- **Lack of transparency**: Both Cambridge Analytica and Facebook were not forthcoming about the extent of the data collection and its uses.

## THERANOS SCANDAL

Theranos, a health technology company founded by Elizabeth Holmes, claimed to have developed revolutionary blood testing technology that could run hundreds of tests using only a few drops of blood.[15,16] The scandal resulted in the dissolution of Theranos, criminal charges against its founders, and potential harm to patients who received inaccurate test results. It also damaged public trust in health technology startups and highlighted the importance of rigorous scientific validation in healthcare innovations. Key failures included:

- **Fraudulent claims**: Theranos made false claims about the capabilities of its technology, which was unable to perform as advertised.
- **Inadequate testing and validation**: The company failed to properly validate its technology or subject it to peer review.
- **Misleading investors and partners**: Theranos provided falsified lab reports and demonstrations to investors and business partners.
- **Violation of clinical laboratory regulations**: The company's labs failed to meet basic quality standards required by regulators.

## BOEING 737 MAX CRISIS

Boeing's 737 MAX aircraft was involved in two fatal crashes in 2018 and 2019, leading to a worldwide grounding of the aircraft model and exposing serious flaws in Boeing's design and certification processes.[17,18] The crisis

13 Barbara Ortutay, Danica Kirka, and Gregory Katz, "Facebook's Zuckerberg Apologizes for 'Major Breach of Trust,'" *AP News*, March 22, 2018, https://apnews.com/article/c8f615be9523421998b4fcc16374ff37.

14 Sara Salinas, "Zuckerberg on Cambridge Analytica: 'We Have a Responsibility to Protect Your Data, and If We Can't Then We Don't Deserve to Serve You,'" *CNBC*, March 21, 2018, https://www.cnbc.com/2018/03/21/zuckerberg-statement-on-cambridge-analytica.html.

15 Gerry Canon, "What the Theranos, Boeing, and Volkswagen Compliance Lapses Have in Common," *ACC Docket*, March 3, 2022, https://www.bc.edu/content/dam/bc1/schools/law/academics/profiles/cv/Caron-Theranos_Boeing_Volkswagen.pdf.

16 U.S. Attorney's Office, Northern District of California, "Theranos Founder Elizabeth Holmes Found Guilty of Investor Fraud," press release, January 4, 2022, https://www.justice.gov/usao-ndca/pr/theranos-founder-elizabeth-holmes-found-guilty-investor-fraud.

17 Canon, "What the Theranos, Boeing, and Volkswagen Compliance Lapses Have in Common."

18 Majority Staff of the Committee on Transportation and Infrastructure, "The Design, Development and Certification of the Boeing 737 Max," September 2020, https://democrats-transportation.house.gov/imo/media/doc/2020.09.15.pdf.

resulted in 346 fatalities, an approximate 20-month grounding of the 737 MAX fleet, billions in financial losses for Boeing and its customer airlines, and a severe blow to Boeing's reputation and the public's trust in aviation safety. It also led to increased scrutiny of aircraft certification processes worldwide. Key failures included:

- **Design flaws**: Boeing implemented the Maneuvering Characteristics Augmentation System (MCAS), a flight control system to counteract the aircraft's tendency to pitch up, without adequately informing pilots or regulators about its functionality.
- **Insufficient safety analysis/inadequate testing and validation**: The company underestimated the control system's potential impact on flight safety.
- **Inadequate training**: Boeing failed to provide comprehensive training on the new control system to pilots.
- **Regulatory oversight lapses:** The FAA delegated too much of the certification process to Boeing itself, compromising the integrity of the safety review process.

# FTX COLLAPSE

FTX, once the third-largest crypto exchange service, collapsed in November 2022 after a surge of customer withdrawals due to reports on its questionable financial valuation practice and close relationship with FTX-affiliated trading firm, Alameda Research.[19] The case garnered significant media coverage due to FTX CEO Sam Bankman-Fried's young age, his previous reputation as an industry leader and philanthropist, and the scale of lost customer assets.[20] The FTX collapse resulted in the loss of billions in customer funds, criminal charges against Bankman-Fried and other executives, and a negative spillover effect on the cryptocurrency market and crypto industry's reputation. The case also exposed the dangerous consequences of corporate misconduct. Key failures included:

- **Poor internal oversight**: FTX did not have a board of directors, or any type of corporate governance structure, leaving Bankman-Fried's activities completely unchecked.
- **Fraudulent practices/misleading users and relevant stakeholders**: FTX established various types of "backdoor" accounts and false reporting mechanisms, misleading shareholders and regulators about the business practices.[21]
- **Misappropriation of customer funds**: Bankman-Fried used customer funds for personal use, investments, and donations.
- **Regulatory evasion**: FTX set up a complex corporate structure to avoid oversight. The organizational chart included over 100 entities (e.g., subsidiaries, affiliates, and interrelated firms) through which Bankman-Fried transferred customer deposits to Alameda Research.[22]

---

19 U.S. Department of Justice Office of Public Affairs, "Samuel Bankman-Fried Sentenced to 25 years for His Orchestration of Multiple Fraudulent Schemes," press release, March 28, 2024, https://www.justice.gov/opa/pr/samuel-bankman-fried-sentenced-25-years-his-orchestration-multiple-fraudulent-schemes.

20 Nemitt Shroff and Cate Reavis, "Sam Bankman-Fried's FTX," MIT Sloan, January 17, 2024, https://mitsloan.mit.edu/sites/default/files/2024-06/.

21 Alexander Osipovich and Angus Berwick, "FTX Employees Found Alameda's Secret Backdoor Months before Collapse," *Wall Street Journal*, October 5, 2023, https://www.wsj.com/finance/ftx-employees-found-alamedas-secret-backdoor-months-before-collapse-7f983fcd.

22 Scott Nover, "The Scrollable, Annotated, Incredibly Complex Org Chart of FTX and Sam Bankman-Fried's Fallen Empire," *Yahoo Tech*, November 17, 2022, https://www.yahoo.com/tech/ftx-bankruptcy-filing-reveals-remarkably-193200722.html.

## CROWDSTRIKE OUTAGE

On July 19, 2024, 8.5 million computers across the globe using the Microsoft Windows operating system experienced the so-called "blue screen of death," an error message on a blue background signifying that a computer is no longer functioning.[23] This incident—since described as the largest IT outage ever—disrupted a multitude of critical infrastructure sectors, including financial services, healthcare, transportation, and emergency services.[24] For example, in the airline industry, the incident forced several major U.S. airlines, including Delta, to ground their planes, which resulted in the cancellation of over 7,000 flights in the span of five days.[25]

While post-mortem reviews remain ongoing, the outage reportedly originated from a faulty update to cybersecurity firm CrowdStrike's Falcon endpoint security product, which was installed on the affected computers. As is the case with nearly all such cybersecurity products, Falcon receives regular updates—often several times each day—to keep abreast of new threat intelligence. In this case, the update included "buggy" code, which caused the affected Windows computers to crash.26 The incident's broad real-world impact across numerous sectors has raised concerns across the public and private sectors about concentration risks to key critical infrastructure services. Key possible failures included:

- **Inadequate risk assessment**: The company failed to foresee the potential impact of a global deployment failure, particularly given their position as a crucial security provider for millions of devices worldwide.
- **Insufficient testing and quality control**: Despite performing automated and manual testing, CrowdStrike appears to have had less rigorous checks for this particular update. Previous successful deployments had given the company confidence in its validation process, which ultimately failed in this instance.[27,28]

# Key Lessons From The Past

The above case studies and identified "key failures" revealed three main categories of compliance failures: Institutional, procedural, and performance failures. These categories are not mutually exclusive. In the AI ecosystem, we expect all three categories of failures to occur, given complex, error-prone and hallucination-prone systems, the growing number of

23 Brian Fung, "We Finally Know What Caused the Global Tech Outage - and How Much It Cost," *CNN*, July 24, 2024, https://www.cnn.com/2024/07/24/tech/crowdstrike-outage-cost-cause/index.html.

24 Lily Hay Newman, "How One Bad CrowdStrike Update Crashed the World's Computers," *WIRED*, July 19, 2024, https://www.wired.com/story/crowdstrike-outage-update-windows/.

25 Michael Liedtke, "CrowdStrike Estimates the Tech Meltdown Caused by Its Bungling Left a $60 Million Dent in Its Sales," *AP News*, August 28, 2024, https://apnews.com/article/crowdstrike-technology-outage-fallout-delta-c287aaded657a1092724b222435c3d16.

26 CrowdStrike, "Falcon Content Update Remediation and Guidance Hub," crowdstrike.com, July 21, 2024, https://www.crowdstrike.com/falcon-content-update-remediation-and-guidance-hub/.

27 "External Technical Root Cause Analysis — Channel File 291," CrowdStrike, August 6, 2024, https://www.crowdstrike.com/wp-content/uploads/2024/08/Channel-File-291-Incident-Root-Cause-Analysis-08.06.2024.pdf.

28 Tom Warren, "CrowdStrike Blames Test Software for Taking down 8.5 Million Windows Machines," *The Verge*, July 24, 2024, https://www.theverge.com/2024/7/24/24205020/crowdstrike-test-software-bug-windows-bsod-issue.

AI incidents, and clear "race to the bottom" dynamics, which signal that safety and security protocols might get overlooked in order to win market advantage.[29]

### Institutional failures

Lack of executive commitment to create a culture of compliance, establish necessary policies, or empower success through the organizational structure (e.g., risk and audit board committees, compliance officer role, quality assurance program), leading to foreseeable failures.

*Examples: Lehman Brothers Bankruptcy, FTX Collapse, Theranos Scandal*

### Procedural failures

Misalignments between an institution's established policies as compared to its internal procedures and staff training required to adhere to those policies.

*Examples: Chernobyl Disaster, Three Mile Island Accident.*

### Performance failures

An employee's failure to follow an established process, or an automated system's failure to perform as intended, leading to an undesirable result.

*Examples: CrowdStrike Outage, Three Mile Island Accident*

These categories also apply in the AI context. For example, deploying insufficiently tested models, obscuring the limitations of AI systems, failing to protect against data misuse or algorithmic bias are potential forms of procedural failures. Institutional failures could include cutting corners to obtain go-to-market advantage, prioritizing production over safety and security. A mix of all three categories of failures could potentially harm individuals or society at large if AI systems malfunction or are misused.

Notably, researchers have extensively studied catastrophic and near-catastrophic failures, exploring conventional engineering approaches to safety in technological systems and the fragile state of nuclear arsenal reliability.[30,31] Across various fields, experts recognize the importance of proactive risk management strategies that integrate human factors, organizational dynamics, and adaptable design approaches to address potential system failures amid complex uncertainties.[32,33] Among the organizations and researchers studying the underlying reasons and implications of system failures, IST has previously dedicated

29 "OECD AI Incidents Monitor (AIM)," OECD.AI, accessed November 11, 2024, https://oecd.ai/en/incidents.

30 Charles Perrow, "Normal Accidents: Living with High Risk Technologies ," www.jstor.org, 1999, https://www.jstor.org/stable/j.ctt7srgf

31 Scott Sagan, The Limits of Safety: Organizations, Accidents, and Nuclear Weapons, (Princeton: Princeton University Press, 2024), https://press.princeton.edu/books/paperback/9780691021010/the-limits-of-safety.

32 Darryl Farber et al., "The Bridge: Linking Engineering and Society," 2020, https://www.nae.edu/File.aspx

33 M.E. Pate-Cornell and J.E. Neu, "Warning Systems and Defense Policy: A Reliability Model for the Command and Control of U.S. Nuclear Forces," *Risk Analysis* 5, no. 2 (June 1985): 121–38, https://doi.org/10.1111/j.1539-6924.1985.tb00160.x.

special attention to examining complex system accidents and their broader security implications, as well as in-depth analysis of failure points for crisis communication systems in the nuclear risk reduction context.[34,35,36]

By drawing parallels between these historic compliance failures and potential issues in the AI context, several key lessons emerge.

- First, the tendency to prioritize rapid innovation and market dominance over safety and ethical considerations, as seen in the Theranos and Boeing cases, is a significant risk in the fast-paced AI industry. AI companies must resist the urge to overstate their capabilities or deploy undertested systems, especially in high-stakes applications like healthcare or autonomous vehicles.
- Second, the lack of transparency and adequate oversight, evident in almost all cases, highlights the need for robust, independent verification of AI systems and their claims. This is particularly crucial given the "black box" nature of many AI algorithms.
- Third, the collection of sensitive data and potential misuse of personal data, as demonstrated by the Cambridge Analytica case study, underscores the critical importance of data governance and privacy protection in AI development and deployment.
- Finally, the regulatory failures observed, especially in the Lehman Brothers and Boeing cases, emphasize the need for adaptive, technologically-informed regulation in the AI space.

The case studies also revealed interesting insights on the trade-offs between open and closed systems, which has become increasingly relevant for AI builders and users.

- In the case of the CrowdStrike outage, which only affected Windows systems, Microsoft's open and flexible approach—which allows trusted third-parties kernel-level access to the Windows operating system—left the system vulnerable to a glitch not of the company's making.
- Meanwhile, the more closed macOS operating system, often referred to as "walled garden," prevents third-party software from having this level of trusted access.
- However, prioritizing security and control comes at the expense of user customization. In the future, AI users will have to make an important decision about picking and deploying a specific type of AI system that keeps these trade-offs in mind.

---

34 Nancy Leveson, "An Engineering Perspective on Avoiding Inadvertent Nuclear War," Institute for Security and Technology, July 2020, https://securityandtechnology.org/wp-content/uploads/2020/07/lleveson_ST_report.pdf.

35 Leah Walker and Alexa Wehsener, "To the Point of Failure: Identifying Failure Points for Crisis Communications Systems," Institute for Security and Technology, January 6, 2023, https://securityandtechnology.org/virtual-library/reports/to-the-point-of-failure-identifying-failure-points-for-crisis-communications-systems/.

36 Leah Walker, "Why We Study Accidents: Complex System Accidents and Their Broader Security Implications," Institute for Security and Technology, February 6, 2023, https://securityandtechnology.org/blog/why-we-study-accidents/.

Another issue that will become increasingly relevant for AI builders and users is the importance of third party evaluators, exemplified by the OceanGate Titan submersible case.

- OceanGate never put the submersible through the standard third-party safety review process, allegedly disregarding warnings about the possible problems.37,38
- Likewise, third-party evaluations of frontier AI models can expose possible procedural and performance failures and avoid catastrophic risks.
- AI builders should therefore engage with credible organizations that have specific expertise in exposing different types of vulnerabilities within their systems.

To mitigate potential compliance failure risks in the AI ecosystem, the AI industry must proactively embrace rigorous testing, transparent reporting of capabilities and limitations, strong data governance practices, and collaborative engagement with regulators to develop effective oversight mechanisms. However, in the absence of AI-focused federal legislation, defining what compliance failure entails, who the relevant actors are, and what the reference points are for safe, secure, and responsible AI development and deployment remains murky.

# Sources of AI Governance

Efforts to govern AI technologies remain nascent but are being rapidly promulgated from a variety of sources, leading to a byzantine patchwork of approaches and requirements. In an attempt to detangle this environment, we introduce five main sources of AI governance: laws & regulations, guidance, norms, standards, and organizational policies.

## Laws & Regulations

AI-related laws include regulations at both the state and federal levels that are already in effect. These laws are either specifically designed to target AI systems, or are broader but could still apply to AI. Many existing laws that address serious concerns like data privacy, safety, bias, information integrity, and general user protection measures do not explicitly mention AI. However, they often still impact AI systems and are expected to create additional compliance needs. For instance, AI system developers and users in the United States can be subject to federal laws such as Equal Employment Opportunity Commission (EEOC) guidelines, the Fair Housing Act (HUD), the Equal Credit Opportunity Act (ECOA), the Age Discrimination

37 Mark Pratt, "How the Unconventional Design of the Titan Sub May Have Destined It for Disaster," *AP News*, June 23, 2023, https://apnews.com/article/titan-titanic-submersible-design-49b8c2a713f316ce5987a394a27d23e8.

38 Aimee Picchi, "Years before Titanic Sub Went Missing, OceanGate Was Warned about 'Catastrophic' Safety Issues," *CBS News*, June 22, 2023, https://www.cbsnews.com/news/missing-titanic-submarine-oceangate-safety-warnings-lawsuits/.

and Employment Act (ADEA), and the Health Insurance Portability and Accountability Act (HIPAA).[39,40,41]

Additionally, the Federal Aviation Administration (FAA), Food and Drug Administration (FDA), Federal Trade Commission (FTC), and the Federal Communications Commission (FCC) have certain regulations that apply to AI systems. These agencies are taking further steps specifically targeting AI systems. For example, the FCC recently declared AI-generated voices—namely, voice cloning technology used in common robocall scams targeting consumers—illegal, and the FTC announced "Operation AI Comply" to address deception and schemes that use or claim to use AI.[42,43]

In the European context, aside from the EU AI Act, the European Parliament's study on the impact of GDPR on AI asserts that even though AI is not explicitly mentioned in the GDPR, many of its provisions are relevant to AI.[44] Industry is aware of this trend; for instance, Microsoft recently published a white paper for customers using their GenAI products that outlined the relevant provisions of GDPR for generative AI.[45]

In the United States, state laws on privacy could also regulate AI systems. Examples include California's Privacy Protection Act, which regulates automated decision-making, Illinois' Biometric Information Privacy Act (BIPA), and California's Fair Employment and Housing Act (FEHA).[46,47]

Separately, new laws have emerged at the state level to address AI systems. For the purposes of this paper, the authors highlight laws relevant to the private sector directed at both AI system builders and users (referred as "deployers") adopted by California, Colorado, and Utah (Table 1). The main compliance requirements under these laws include transparency, internal

39 Delaram Rezaeikhonakdar, "AI Chatbots and Challenges of HIPAA Compliance for AI Developers and Vendors," *Journal of Law, Medicine & Ethics* 51, no. 4 (January 1, 2023): 988–95, https://doi.org/10.1017/jme.2024.15.

40 "What Is the EEOC's Role in AI?" US Equal Employment Opportunity Commission, accessed November 18, 2024, https://www.eeoc.gov/sites/default/files/2024-04/20240429.

41 Courtney Dankworth et al., "Adverse Action Notice Compliance Considerations for Creditors That Use AI," *Business Law Today*, American Bar Association, October 30, 2023, https://www.americanbar.org/groups/business_law/resources/business-law-today/2023-november/adverse-action-notice-compliance-considerations-for-creditors-that-use-ai.

42 U.S. Federal Communications Commission, "FCC Makes AI-Generated Voices in Robocalls Illegal," press release, February 8, 2024, https://www.fcc.gov/document/fcc-makes-ai-generated-voices-robocalls-illegal.

43 Julia Solomon Ensor, "Operation AI Comply: Continuing the Crackdown on Overpromises and AI-Related Lies," *Business Blog* (blog), U.S. Federal Trade Commission, September 25, 2024, https://www.ftc.gov/business-guidance/blog/2024/09/operation-ai-comply-continuing-crackdown-overpromises-ai-related-lies.

44 "The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence," European Parliamentary Research Service, June 2020, https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf.

45 Manny Sahota, "Introducing Our New Whitepaper: GDPR & Generative AI – a Guide for Customers," *Microsoft Community Hub* (blog), Microsoft, June 4, 2024, https://techcommunity.microsoft.com/t5/security-compliance-and-identity/introducing-our-new-whitepaper-gdpr-amp-generative-ai-a-guide/ba-p/4158935.

46 White & Case LLP, "AI Watch: Global Regulatory Tracker - United States," www.whitecase.com, May 13, 2024, https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-states.

47 "DEREK MOBLEY v. WORKDAY INC," Findlaw, 2024, https://caselaw.findlaw.com/court/us-dis-crt-n-d-cal/116378658.html.

assessments, training, incident reporting, nondiscrimination, and providing users the ability to "opt out."

### Table 1: State Bills Signed Into Laws in 2024 Targeting AI Developers and Users

**UTAH**

**SB 149: Artificial Intelligence Policy Act** *Effective 5/1/2024* — $2,500 - 5,000 per violation

➔ Amends the Utah consumer protection and privacy laws to require disclosure, in certain circumstances, of artificial intelligence (AI) use to consumers.

**Scope:** Deployers of any commercial communication using "generative AI" (Defined as a system that is trained on data; interacts with a person using text, audio, or visual communication; and generates non-scripted outputs similar to outputs created by a human, with limited or no human oversight.)

**CALIFORNIA**

**AB 2013: Generative Artificial Intelligence: Training Data Transparency** *Effective 1/1/2026* — Not specified

➔ Imposes new disclosure requirements on the developers of generative artificial intelligence (GenAI) systems and services that are made available to Californians.

**Scope:** Developers using "generative AI" (Defined as "artificial intelligence that can generate derived synthetic content, such as text, images, video, and audio, that emulates the structure and characteristics of the artificial intelligence's training data.")

**SB 942 California AI Transparency Act** *Effective 1/1/2026* — $5,000 per violation

➔ Requires making certain AI detection tools available at no cost to users.

**Scope:** "Covered providers" are defined as a person that creates, codes, or otherwise produces a generative artificial intelligence system that has over 1,000,000 monthly visitors or users and is publicly accessible within the geographic boundaries of the state.

**COLORADO**

**SB 205: Concerning Consumer Protections in Interactions with Artificial Intelligence Systems** *Effective 2/1/2026* — $20,000 per violation

➔ Requires a developer of a high-risk artificial intelligence system to use reasonable care to protect consumers from any known or reasonably foreseeable risks of algorithmic discrimination in the high-risk system.

**Scope:** Developers and deployers of "high-risk" AI systems (defined as AI systems that make or significantly influence "consequential decisions" in areas such as employment, housing, credit, education, and healthcare).

# Guidance

In the AI context, guidance refers to non-legally binding recommendations that carry significant weight, and therefore typically influence AI governance and compliance. Guidance is usually issued by governments and their respective agencies. Governments around the world are also adopting AI strategies, governance frameworks, and other documents that, while not laws, serve as important guidance for builders and users considering operating in those countries.[48] Examples include the National Institute of Standards and Technology's (NIST) AI Management Framework, the President Biden's National Security Memorandum on AI, the Framework to Advance AI Governance and Risk Management in National Security, and

48 IAPP, "Global AI Law and Policy Tracker," IAPP, last updated October 2024, accessed November 11 2024, https://iapp.org/media/pdf/resource_center/global_ai_law_policy_tracker.pdf.

other relevant frameworks published in accordance with the Executive Order 14110 on the Safe, Secure and Trustworthy Development and Use of AI.

# Norms

We define norms in the context of AI as evolving principles that guide behavior and set expectations regarding the development and use of AI technologies. Unlike guidance, norms are often designed by multilateral organizations or arise from international convenings. These norms reflect societal values and ethical considerations, influencing how AI is perceived and governed. They are typically proposed by regional, multinational, and international organizations. Examples include the Organisation for Economic Co-operation and Development (OECD) principles on AI, the United Nations Educational, Scientific and Cultural Organization's (UNESCO's) Recommendations on AI Ethics, the Hiroshima Process, the Seoul Declaration, the Bletchley Declaration, UN Global Digital Compact, and other AI governance frameworks introduced by the United Nations and its organizations and bodies.[49,50]

# Standards

We refer to AI standards as those developed within a standards development organization (SDO) to govern AI systems, either specifically for AI developers and users or that are broadly applicable to them. For instance, the Institute of Electrical and Electronics Engineers (IEEE) standards 7000 and 7002, while not AI-specific, can still be useful for AI developers and users. The number of standards related to AI has significantly increased from 14 in 2020 to 117 as of June 2024.[51] The main SDOs currently developing AI-specific standards include the International Organization for Standardization (ISO), IEEE, and the UN's International Telecommunication Union (ITU).

# Organizational Policies

Organizational policies set out internal oversight and accountability procedures and practices. Organizational policies are often confidential and publicly unavailable. Many of the policies develop to adhere to already established laws and some AI-focused norms, like the Hiroshima Process and Seoul Declaration, can inspire these internal policies. Additionally, several AI labs

49 UN Secretary General's High-level Advisory Body on AI, "Governing AI for Humanity," United Nations, September 19, 2024, https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf.

50 Office of the Secretary-General's Envoy on Technology, "Global Digital Compact," United Nations, September 22, 2024, https://www.un.org/techenvoy/global-digital-compact.

51 UN Secretary General's High-level Advisory Body on AI, "Governing AI for Humanity."

rely on voluntary commitments, such as those signed in in September 2023 during a White House meeting on the topic.[52,53] For instance, Anthropic publishes its "Voluntary Commitments Tracker" highlighting progress on its voluntary commitments.[54] Other organizational policies include Anthropic's Responsible Scaling Policy, Google's Secure AI Framework, and OpenAI's safety practices.[55,56,57]

# Defining Compliance Failure in the AI Context

In the absence of comprehensive, AI-focused federal legislation in the United States, we define compliance failure in the AI ecosystem as the failure to align with existing laws, government-issued guidance, globally accepted norms, standards, voluntary commitments, and organizational policies (whether publicly announced or confidential) that focus on responsible AI governance. This concept encompasses a spectrum of potential infractions, ranging from oversights with limited repercussions to severe violations that could lead to far-reaching consequences. In the rapidly evolving field of AI, compliance failure presents unique challenges due to the complex interplay between cutting-edge technology, regulatory frameworks, and societal expectations. Understanding the nature and scope of compliance failure in AI is crucial for the responsible development and deployment of these powerful technologies. Even the most ardent proponents of AI advancement recognize the necessity of mitigating potential risks and the importance of maintaining public trust.

The challenge of mitigating compliance failure in the AI ecosystem is threefold, as it involves the actions and behaviors of three separate classes of actors: regulators, builders, and users. Regulators encompass any entity that creates compliance mechanisms, such as professional boards, governments, or other actors who establish guidelines and regulations for AI systems.

52 "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI," The White House, July 21, 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

53 The White House, "Voluntary AI Commitments," September 2023, https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.

54 "Tracking Voluntary Commitments," Anthropic, last updated November 18, 2024, https://www.anthropic.com/voluntary-commitments.

55 "Google's Secure AI Framework (SAIF)," Google, accessed November 14, 2024, https://safety.google/intl/en_us/cybersecurity-advancements/saif/.

56 "Anthropic's Responsible Scaling Policy," Anthropic, September 19, 2023, https://www.anthropic.com/news/anthropics-responsible-scaling-policy.

57 OpenAI, "Safety & Responsibility," openai.com, 2024, https://openai.com/safety/.

Builders refer to the individuals or organizations responsible for developing the models including AI labs, startups, and tech companies. Users include all other entities who deploy or utilize the technology, including enterprises integrating AI systems into their services and internal operations.

Compliance failure, therefore, arises from a disconnect or breakdown in the relationship among these three actors. For example, regulators may create overly strict or impractical guidelines, leading builders to ignore them or devise workarounds. Alternatively, builders may adhere to compliance principles, but users could circumvent these structures, leading to unintended outcomes. In some cases, users may utilize models in alignment with compliance mechanisms, but builders may have failed to implement them robustly, rendering them ineffective. Furthermore, even when builders and users follow compliance mechanisms, regulators might not fully understand workflows, user behaviors, or system capabilities, leading to an inability to prevent adverse outcomes.

Regulated industries usually have clear compliance points of reference, such as regulatory bodies, regulations, industry-specific laws, and standards. This is not yet the case with the AI ecosystem in the United States.

# Current State of Play

Ongoing litigation against companies deploying AI systems highlights the importance of proactively implementing the measures for safe, secure, and responsible AI development. Recent court cases involving companies like iTutor, Clearview AI, HireVue, and Workday (further explored below) teach a crucial lesson: compliance with privacy, anti-discrimination, and transparency laws must be foundational, not an afterthought.

## Data Privacy & User Consent

The first trend we observed in AI compliance revolves around data privacy, data processing, and proper user consent forms. Clearview AI, a facial recognition company, faced fines and public backlash over privacy concerns related to its facial recognition technology. In 2020, The New York Times reported that Clearview AI was building tracking and surveillance tools using biometric identifiers, capturing more than three billion faceprints.[58] A lawsuit filed against the company in the same year alleged violation of Illinois residents' privacy rights under BIPA.[59]

58 Kashmir Hill, "The Secretive Company That Might End Privacy as We Know It," *The New York Times*, January 18, 2020, https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

59 "Court Cases: ACLU v. Clearview AI," American Civil Liberties Union, May 11, 2022, https://www.aclu.org/cases/aclu-v-clearview-ai.

The parties reached a settlement in the United States, but Clearview AI faced fines in Europe for violating GDPR, estimated at $110 million, including for unlawful processing of personal data.[60,61,62]

Similarly, HireVue, a software vendor that conducts video- and game-based "pre-employment" assessments using facial recognition technology and proprietary algorithms, faced allegations of improperly collecting and using biometric data.[63] In November 2019, the Electronic Privacy Information Center (EPIC) filed a complaint urging the FTC to investigate HireVue's business practices, saying the company's use of AI systems that scan people's faces and voices constituted a pervasive threat to American workers. Notably, the complaint cited OECD AI principles, saying that it failed to comply with the minimum standards for AI-based decision making set out in the principles.[64] In response, HireVue announced in 2021 that it would stop relying on "facial analysis."[65] However, a new lawsuit emerged in 2022, claiming that the company violated BIPA by improperly collecting and using biometric data.[66]

The Clearview AI and HireVue cases underscore the importance of proper data privacy mechanisms, including user consent forms and transparency in data processing practices. Implementing these practices helps companies maintain their reputation, protect end-user privacy, and avoid financial liabilities.

# Algorithmic Bias

The second emerging trend involves alleged algorithmic biases, as seen in the iTutor and Workday cases. iTutor Group (composed of iTutorGroup, Inc.; Shanghai Ping'An Intelligent Education Technology Co., Ltd.; and Tutor Group Limited) is an online platform for hiring U.S.- based tutors in China. In September 2023, iTutor settled an EEOC lawsuit accusing the company of programming its application software to automatically reject female applicants

60 Adrianne Appel, "Clearview AI's GDPR Fines Rise to $110M Total after Latest Penalty by Dutch DPA," *Compliance Week*, September 9, 2024, https://www.complianceweek.com/regulatory-enforcement/clearview-ais-gdpr-fines-rise-to-110m-total-after-latest-penalty-by-dutch-dpa/35338.article.

61 "The French SA Fines Clearview AI EUR 20 Million," European Data Protection Board, October 20, 2022, https://www.edpb.europa.eu/news/national-news/2022/french-sa-fines-clearview-ai-eur-20-million_en.

62 "The French SA Fines Clearview AI EUR 20 Million."

63 "In Re HireVue Consumer Cases," Electronic Privacy Information Center, accessed November 1, 2024, https://epic.org/documents/in-re-hirevue/.

64 Drew Harwell, "Rights Group Files Federal Complaint against AI-Hiring Firm HireVue, Citing 'Unfair and Deceptive' Practices," *The Washington Post*, November 6, 2019, https://www.washingtonpost.com/technology/2019/11/06/prominent-rights-group-files-federal-complaint-against-ai-hiring-firm-hirevue-citing-unfair-deceptive-practices/.

65 Will Knight, "Job Screening Service Halts Facial Analysis of Applicants," *WIRED*, January 12, 2021, https://www.wired.com/story/job-screening-service-halts-facial-analysis-applicants/.

66 Adam Forman, Alexander Franchili, and Naomi Friedman, "Deyerler v. HireVue Expands Biometric Privacy Law to AI Video Interview Platform," *Workforce Bulletin* (blog), Epstein Becker Green, March 6, 2024, https://www.workforcebulletin.com/deyerler-v-hirevue-expands-biometric-privacy-law-to-ai-video-interview-platform.

aged 55 or older and male applicants aged 60 or older. The software automatically rejected more than 200 applicants, a practice that constituted discriminatory hiring practices under the Age Discrimination and Employment Act.[67]

Similarly, Workday, a cloud-based software platform for managing business finances and human resources, faced legal scrutiny in the case of Mobley v. Workday, Inc. A job applicant alleged that the company's AI screening tools discriminated against him, violating federal and California employment laws.[68] While the case is still ongoing, court proceedings document an interesting statement on software vs human decision makers. The judge stated that, "[n]othing in the language of the federal anti-discrimination statutes or the case law interpreting those statutes distinguishes between delegating functions to an automated agent versus a live human one." The judge continued, "Drawing an artificial distinction between software decision makers and human decision makers would potentially gut anti-discrimination laws in the modern era."[69] The outcome of this case will be highly consequential for similar companies using AI tools for hiring.

Both cases highlight the importance of carefully designing and deploying AI-based software to minimize the risk of algorithmic bias.

Historical and modern case studies demonstrate the possibility of more severe consequences if companies fail to comply in high-risk environments. For example, if an AI system designed for use in healthcare settings failed to comply with privacy regulations, it might lead to a data breach, exposing sensitive patient information. As AI continues to penetrate various business operations, from hiring practices to consumer applications, companies must take steps to ensure transparency and accountability.

The long-term effects of AI compliance failures could shape the regulatory landscape itself. If compliance failures result in significant negative consequences, policymakers may feel compelled to introduce more stringent regulations, potentially stifling innovation and limiting the development of beneficial AI applications. On the other hand, if compliance failures are not adequately addressed, the lack of effective regulation could lead to a "wild west" scenario, where the absence of proper oversight enables the proliferation of harmful AI practices.

67 "ITutorGroup to Pay $365,000 to Settle EEOC Discriminatory Hiring Suit," *Newsroom* (blog), U.S. Equal Employment Opportunity Commission, September 11, 2023, https://www.eeoc.gov/newsroom/itutorgroup-pay-365000-settle-eeoc-discriminatory-hiring-suit.

68 Mobley v. Workday, Inc., No. 23-cv-00770-RFL, 2024 U.S. LEXIS 126336 (N.D. Ca. 7/12/24), https://caselaw.findlaw.com/court/us-dis-crt-n-d-cal/116378658.html.

69 Mobley v. Workday, Inc.

# Outlook

The AI ecosystem faces unique challenges in maintaining compliance due to its rapidly evolving nature, complexity of the AI systems, and the lack of established definitions of safety.

### Ambiguous AI safety definitions and the rapid pace of development challenges governance and, potentially, AI adoption across regulated industries.

Finding the acceptable risk threshold for each business and use case will be challenging, as the developments from AI—and their associated implications—will be difficult to predict and manage. Well-regulated industries that utilize cutting-edge technology (i.e. financial services, healthcare) typically have clear definitions of safety and reliability. In comparison, AI's notion of "safety" is more ambiguous, leaving it up to interpretation and making it difficult to articulate clear safety regulations.[70] This ambiguity is an increasing concern given the rapid pace of AI development, which often outstrips the pace of regulatory efforts in this ecosystem. As a result, the deployment of AI models could occur in a gray area, in which regulatory frameworks may not fully or clearly address emerging risk scenarios brought about by AI innovations, which might disincentivize adoption.

### Interpretability challenges will hinder development of compliance mechanisms.

The complexity and opacity of AI models also contributes to potential compliance issues. Policymakers, regulators, and the general public may lack sufficient understanding of the technical risks and opportunities associated with AI models. The deepening literacy silos between developers and regulators adds a layer of complexity to responsible use and informed decision-making. The interpretability challenges posed by these "black box" systems make it difficult for regulators to design effective and long-lasting compliance mechanisms. This lack of understanding can lead to fear-driven regulations that may hinder growth and development in the field. While leading labs are working toward the improvement of model interpretability, there remain unanswered questions surrounding opacity that challenge regulatory compliance.[71]

70 Brian Judge, Mark Nitzberg, and Stuart Russell, "When Code Isn't Law: Rethinking Regulation for Artificial Intelligence," *Policy and Society*, 2024, 00(00), 1–13, DOI: https://doi.org/10.1093/polsoc/puae020.

71 Adly Templeton, et al., "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet," Transformer Circuits Thread, Anthropic, 2024, https://transformer-circuits.pub/2024/scaling-monosemanticity/.

### AI Agents will blur the lines of liability in the automated world.

The concept of AI systems that autonomously perform tasks, known as AI "agents," gained renewed attention due to advancements in large language models.[72] These agents are expected to transform society and the economy; while they currently remain rudimentary, they can also become a source of serious harm.[73] AI agents could become the subject of existing laws and some thought leaders have already started to discuss the difficulties of tracing an AI agent back to the human who initially developed and deployed it, and thus assigning liability.[74] The proliferation of AI agents and the rise of multiagent environments can create feedback loops where decisions based on past data influence future outcomes, and any causal connection between the original deployer's intent and later outcomes will inevitably attenuate. This scenario could enhance and reinforce biases or inaccuracies, or worse yet, leave the human altogether out of the loop.

# Conclusion

The implications of compliance failure in the AI ecosystem can vary significantly, ranging from minor infractions with minimal consequences to catastrophic events that have far-reaching effects on society and technology usage. As AI systems become increasingly integrated with the global economy, the potential effects of compliance failures extend far beyond the AI industry itself. AI, by its very nature, is cross-sectoral. As such, it is and will continue to revolutionize sectors such as manufacturing, finance, real estate, healthcare, and public safety. Consequently, compliance failures within AI labs could have an outsized impact across the broader economy and society. In addition, failure to enforce AI compliance might also generate a dangerous precedent to "build quickly, secure later"—one that, as we have repeatedly experienced is not a winning approach.

So, what exactly should AI builders and users do to avoid or minimize compliance failure risks? What are the business incentives for proactively developing compliance mechanisms and practices? Part 2 of this report, slated for publication in early 2025, will offer a list of targeted, actionable strategies for mitigating compliance risks using the AI Lifecycle Framework and will present various potential opportunities for the Return on Investment (ROI) for AI builders and users.

---

72 Iason Gabriel et al. "The Ethics of Advanced AI Assistants," arXiv:2404.16244, 2024, https://arxiv.org/pdf/2404.16244.

73 Steve Kelly, Jennifer Tang, and Tiffany Saade, "The Implications of Artificial Intelligence in Cybersecurity: Shifting the Offense Defense Balance," Institute for Security and Technology, October 2024, https://securityandtechnology.org/wp-content/uploads/2024/10/The-Implications-of-Artificial-Intelligence-in-Cybersecurity.pdf.

74 Gillian Hadfield, "How to Prevent Millions of Invisible Law-Free AI Agents Casually Wreaking Economic Havoc," *Fortune*, October 17, 2024, https://fortune.com/2024/10/17/ai-agents-law-economy/.

# Appendix

## Case Studies in AI-Adjacent Industries

### Three Mile Island Accident, 1979

**nuclear**

*"A combination of personnel errors, design deficiencies, and component failures caused the TMI accident."*[75]

**Key failures**: Design flaws, equipment malfunction, human errors, communication breakdowns.[76]

- Partial nuclear meltdown
- Increase in public fear and mistrust toward nuclear energy
- Estimated $973 million in cleanup costs[77]

### Chernobyl Disaster, 1986

**nuclear**

*"The accident arose due to a deficient safety culture, such as a positive reactivity coefficient and a flawed shut down system, which had been known but not corrected."*[78]

**Key failures**: Flawed reactor design, violation of safety procedures, inadequate safety culture, poor emergency response protocol.

- Catastrophic nuclear disaster
- Massive environmental contamination
- Hit to local economy, primarily agriculture
- Thousands of casualties from the initial explosion and serious radiation illnesses[79]

### Space Shuttle Challenger Disaster, 1986

**aerospace**

*The failure occurred because of a flawed design that was excessively sensitive to several factors, including temperature. Those who made the decision to launch were unaware of the recent history of problems with the O-ring. "The unrelenting pressure to meet the demands of an accelerating flight schedule might have been adequately handled by NASA if it had insisted upon the exactingly thorough procedures that were its hallmark during the Apollo program."*[80]

**Key failures**: Design flaws, disregarding warnings from the engineers, overlooking safety procedures, management pressure to meet the program goals

- Loss of seven crew members
- Program suspension
- Erosion of public confidence in the space program

**Key:** Institutional Failures | Performance Failures | Procedural Failures

75 "Backgrounder On Three Mile Island Accident," United States Nuclear Regulatory Commission, fact sheet, last updated March 28, 2024, https://www.nrc.gov/reading-rm/doc-collections/fact-sheets/3mile-isle.html.

76 "Three Mile Island Accident," World Nuclear Association, information library, last updated October 11, 2022, https://world-nuclear.org/information-library/safety-and-security/safety-of-plants/three-mile-island-accident.

77 "Three Mile Island Accident."

78 International Nuclear Safety Advisory Group, "The Chernobyl Accident: Updating of INSAG-1," INSAG Series 7, International Atomic Energy Agency, 1993, https://www-pub.iaea.org/MTCD/publications/PDF/Pub913e_web.pdf.

79 United Nations Department of Media Relations, "Chernobyl: The True Scale of the Accident," press release, DEV/2539, June 9, 2005, https://press.un.org/en/2005/dev2539.doc.htm.

80 William Rogers, "Report to the President by the Presidential Commission on the Space Shuttle Challenger Accident," 1986, https://sma.nasa.gov/SignificantIncidents/assets/rogers_commission_report.pdf.

## Lehman Brothers Bankruptcy, 2008

**financial services**

*"Lehman had an aggressive CEO in Dick Fuld; a "countercyclical" growth strategy; a firm culture that rewarded risk; questionable accounting policies; never before seen levels of market volatility; and a high leverage business model employed by many investment banks."*[81]

**Key failures**: Excessive leverage, fraudulent accounting, inadequate risk management, regulatory oversight lapses

- Widespread economic downturn
- Loss of tens of thousands of jobs
- Erosion of public trust in financial institutions

## Cambridge Analytica Scandal, 2018

**social media**

*The FTC alleged that Cambridge Analytica used deceptive practices involving the collection of personal information from Facebook users for targeted political and commercial advertising through an application on the Facebook platform called the "GSRApp," also known publicly as the "thisisyourdigitallife" app.*[82]

**Key failures**: Unauthorized data collection, misuse of personal information, inadequate data protection measures, lack of transparency

- Increased scrutiny of data privacy
- Erosion of public trust
- Bankruptcy of Cambridge Analytica
- Financial burdens ($5 billion civil penalty)[83]

## Theranos Scandal, 2018

**biotech**

*Theranos claimed to have developed revolutionary blood testing technology that could run hundreds of tests using only a few drops of blood. Theranos made false claims about the capabilities of its technology, which was unable to perform as advertised.*[84]

**Key failures**: Fraudulent claims, inadequate testing and validation, misleading investors and partners, violation of clinical laboratory regulations

- Company dissolution
- Criminal charges against founders
- Potential harm to patients
- Damaged trust in health tech startups

**Key:** Institutional Failures  Performance Failures  Procedural Failures

81 Stuart C. Gilson, Kristin Mugford, and Sarah L. Abbott, "The Rise and Fall of Lehman Brothers," www.hbs.edu, January 2017 (revised January 2019), https://www.hbs.edu/faculty/Pages/item.aspx?num=52147.

82 Joseph Simons et al., "82 3107 United States of America before the Federal Trade Commission Commissioners," 2019, https://www.ftc.gov/system/files/documents/cases/182_3107_cambridge_analytica_administrative_complaint_7-24-19.pdf.

83 U.S. Department of Justice Office of Public Affairs, "Facebook Agrees to Pay $5 Billion and Implement Robust New Protections of User Information in Settlement of Data-Privacy Claims," press release, July 23, 2019, https://www.justice.gov/opa/pr/facebook-agrees-pay-5-billion-and-implement-robust-new-protections-user-information.

84 U.S. Attorney's Office, Northern District of California, "Theranos Founder Elizabeth Holmes Found Guilty of Investor Fraud," press release, January 4, 2022, https://www.justice.gov/usao-ndca/pr/theranos-founder-elizabeth-holmes-found-guilty-investor-fraud.

## Boeing 737 MAX Crisis, 2018-2019

aerospace

*"Technical design flaws, faulty assumptions about pilot responses, and management failures by both The Boeing Company (Boeing) and the Federal Aviation Administration (FAA) played instrumental and causative roles in the chain of errors that led to the crashes."*[85]

**Key failures**: Design flaws, insufficient safety analysis, inadequate training, regulatory oversight lapses

- 346 fatalities
- Worldwide fleet grounding
- Severe reputational damage

## FTX Collapse, 2022

financial services

*Bankman-Fried diverted billions in customer deposits from FTX to Alameda, using the funds for personal investments, political donations, and real estate. He employed various fraudulent tactics to allow Alameda unlimited withdrawals, making false statements to financial institutions, inflating FTX's financials to investors, and backdating documents to cover up his misconduct.*[86]

**Key failures**: Poor internal oversight, fraudulent practices, misappropriation of customer funds, regulatory evasion

- Billions lost in customer funds
- Criminal charges against executives
- Negative impact on crypto market
- Industry-wide reputational damage

# Ongoing Inquiries

## Titan Submersible Implosion (ongoing investigation)[87], 2022

maritime

*Testimonies highlight "manufacturing defects and problems following an earlier dive and reveals that OceanGate conducted no testing or remedial work despite concerns with the hull";*[88] *"Titan was rebuilt with a new hull that was never tested to industry norms nor certified by an independent third-party agency."*[89]

(possibly)

**Key failures**: Design flaws, absence of third-party certification

- Loss of five lives
- Investigation into deep-sea tourism
- Industry-wide reputational damage

**Key:**  Institutional Failures  Performance Failures  Procedural Failures

85 Majority Staff of the Committee on Transportation and Infrastructure, "The Design, Development and Certification of the Boeing 737 Max," September 2020, https://democrats-transportation.house.gov/imo/media/doc/2020.09.15.pdf.

86 U.S. Department of Justice Office of Public Affairs, "Samuel Bankman-Fried Sentenced to 25 Years for His Orchestration of Multiple Fraudulent Schemes," press release, March 28, 2024, https://www.justice.gov/opa/pr/samuel-bankman-fried-sentenced-25-years-his-orchestration-multiple-fraudulent-schemes.

87 United States Coast Guard, "Titan Submersible - Coast Guard Marine Board of Investigation," www.news.uscg.mil, last updated September 23, 2024, https://www.news.uscg.mil/News-by-Region/Headquarters/Titan-Submersible/.

88 Mark Harris, "Titan Submersible Hearings Spotlight Multiple Issues with Its Carbon Fiber Hull," *WIRED*, September 25, 2024, https://www.wired.com/story/titan-submersible-hearings-spotlight-multiple-issues-with-its-carbon-fiber-hull/.

89 "Titan Submersible Hearings Spotlight Multiple Issues."

## 23andMe Data Breach, 2023

**biotech**

*Even though the company admitted to no wrongdoing as part of the settlement, the data breach still exposed the shortcomings that led to undetected cyberattacks and exposure of customer data.[90] The breach exposed millions of user profiles and their health data, ancestry, family connections, and other types of sensitive information. Some of the data, including names, addresses, and genetic heritage, ended up on dark web forums.[91]*

(possibly)

**Key failures**: Inadequate risk assessment, inadequate internal oversight, regulatory oversight lapses

- Reputational damage
- Loss of customer trust
- Regulatory investigations

## CrowdStrike Outage, 2024

**cybersecurity**

*"The outage was caused by a defect found in a Falcon content update for Windows hosts."[92]*

(possibly)

**Key failures**: Error in the system, inadequate risk assessment, insufficient testing and quality control

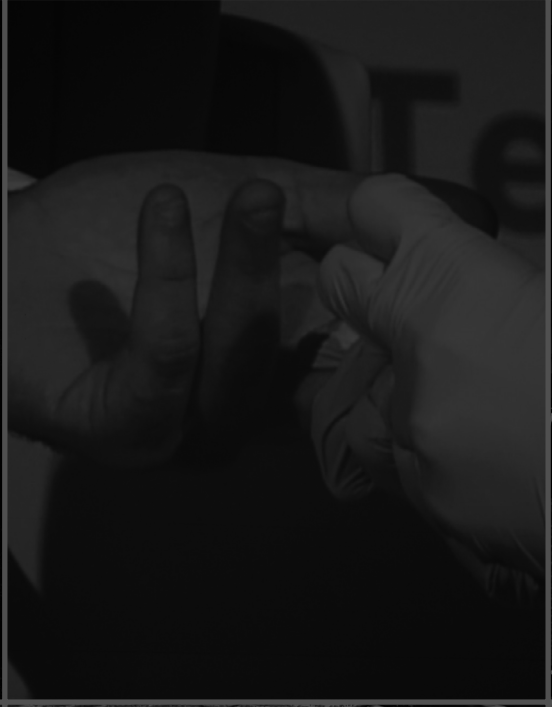
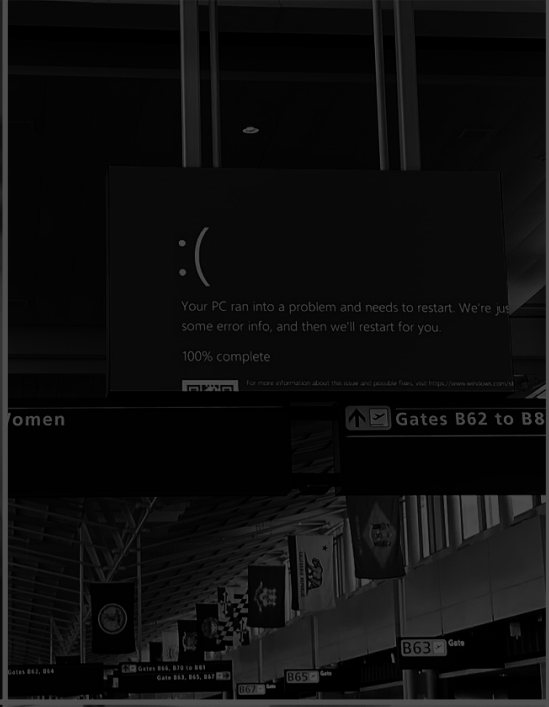
- Disrupted critical IT services across multiple sectors
- Billions lost as a result of service delays

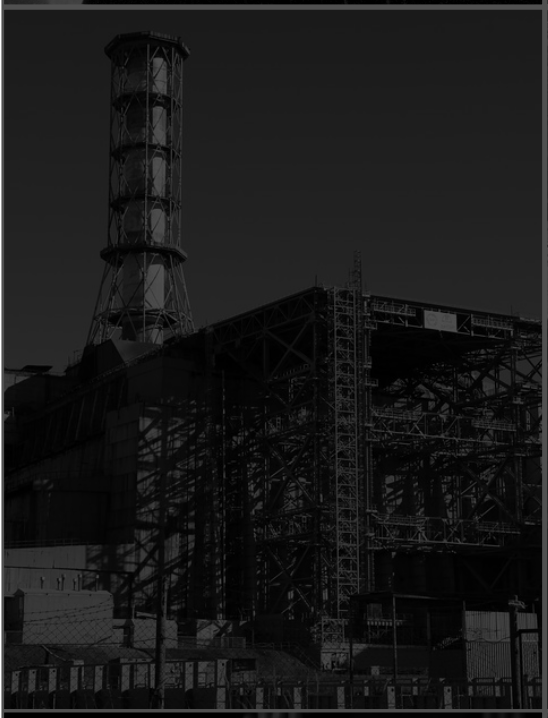
**Key:** Institutional Failures  Performance Failures  Procedural Failures

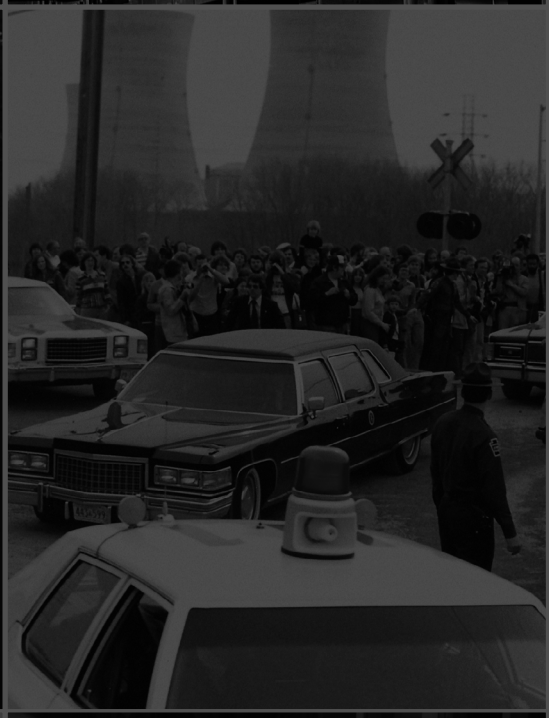
90 Lorenzo Franceschi-Bicchierai, "23andMe Admits It Didn't Detect Cyberattacks for Months," *TechCrunch*, January 25, 2024, https://techcrunch.com/2024/01/25/23andme-admits-it-didnt-detect-cyberattacks-for-months/.

91 Edward Helmore, "Genetic Testing Firm 23andMe Admits Hackers Accessed DNA Data of 7m Users," *The Guardian*, December 5, 2023, https://www.theguardian.com/technology/2023/dec/05/23andme-hack-data-breach.

92 George Kurtz, "To Our Customers and Partners," *CrowdStrike* (blog), July 19, 2024, https://www.crowdstrike.com/en-us/blog/to-our-customers-and-partners/.






# INSTITUTE FOR SECURITY AND TECHNOLOGY

www.securityandtechnology.org

info@securityandtechnology.org