Policy Brief

Operationalizing AI Guidance

A Reference Guide for Translating High-Level Goals into Practical Implementation

Authors
Kyle Crichton
Abhiram Reddy
Jessica Ji

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

April 2026

Executive Summary

Organizations continue to struggle with successfully integrating artificial intelligence into their operations, leading to cases where the technology fails to support, or even undermines, the organization's mission. Researchers in academia, industry, and civil society have done tremendous work in developing techniques that can assist practitioners in their adoption of AI technology. Yet these contributions can easily be lost in the sea of collective AI efforts. While recent work to synthesize this research in key areas has begun, there remains no comprehensive guide to help practitioners navigate the maze of information. To address this gap, CSET researchers have collated over 1,200 resources from the breadth of academic papers, industry reports, and grey literature with the aim of making AI safety, security, and governance practices more accessible and easier to integrate into an organization's existing processes. Building on CSET's harmonized AI framework, this report collates actionable guidance from these resources into an easy-to-navigate format and connects the information back to core recommended practices, establishing a critical link between high-level principles and practical implementation details. This report aims to answer the core questions that practitioners need to ask to successfully operationalize AI best practices.

What: We identify the key techniques, tools, and practices that organizations should implement in adopting AI systems based on an extensive review of government guidelines, industry reports, academic papers, and the broader set of grey literature.

When: We map these implementation steps across the process of AI adoption, including planning, development, deployment, operation, and improvement.

Where: We provide guidance at all levels of the organization, ranging from organization-wide operations and practices to granular details at the data, model, and system level.

Who: We identify the functional teams within the organization that are most likely involved in carrying out each recommended implementation step.

How: We provide an extensive set of references alongside the implementation steps, identifying valuable resources from the staggering volume of available information.

Why: We link each step in our guide back to the recommendations from the harmonized framework that it serves to implement.

In addressing these key components, this report serves as a practical reference guide that bridges the gap between principles and practice for those adopting AI technology.

Table of Contents

Introduction

Organizations today face mounting pressure to adopt artificial intelligence technology, yet many struggle to successfully integrate AI into their structure and operations such that it supports the organization's mission and furthers its goals [1]. On the one hand, a variety of existing reports provide principles and frameworks intended to inform organizations of **what** they should do when adopting AI systems. On the other, there is an extensive body of industry white papers, academic research, and grey literature that provide details on **how** to implement a specific technique or practice related to AI. While there is value in both high-level abstraction and granular detail, organizations need guidance that aligns the two, mapping out what needs to be done and providing the resources on how to do it [2]. This information can be difficult for organizations to find on their own, crowded out by a deluge of low-quality search results, blog posts, and articles on the internet—a trend that has only worsened with the flood of AI-generated content [3, 4, 5]. Several reports—including the National Institute of Standards and Technology's (NIST) AI Risk Management Framework Playbook, several guides from the Open Worldwide Application Security Project (OWASP) Foundation, and the Partnership on AI's governance toolkit—have begun to address this need, tying together high-level practices with practical implementation details [6, 7, 8, 9]. Yet, each only covers a piece of the overall picture, leaving practitioners without an overarching road map for responsible AI adoption.

This report aims to address that problem by providing an extensive reference guide that (1) covers the range of AI safety, security, and governance practices and (2) ties recommendations for **what** practices organizations should implement to **how** they should do so, including **when**, **where**, **why**, and **by whom**. We build on previous CSET research that condensed the collective knowledge of 52 AI guidance documents into a single harmonized framework [10]. Where that first report lays out a holistic picture of AI best practices, this research provides the link between those high-level recommendations and the detailed reports that illustrate how to implement them. We have waded through the breadth of government and industry reports, academic papers, and grey literature to identify valuable sources that practitioners rely on. Making AI guidance more tangible, this report provides answers to the fundamental questions organizations have when implementing AI best practices:

- **What should I do?** Based on our research, we provide a series of practical implementation steps derived from CSET's harmonized AI framework.
- **When?** We organize these steps according to the stage in the AI adoption life cycle (e.g., planning, development, deployment) in which they occur for easy and logical reference.

- **Where?** We structure the recommendations within each stage according to the level of implementation (e.g., operations, data, model, system) at which these practices occur within the organization.
- **By whom?** We identify the functional teams within the organization that the content is most applicable for, while trying not to assume certain organizational structures or prescribe specific roles and responsibilities.
- **How?** We provide an extensive set of references alongside the implementation steps, identifying valuable resources from the staggering volume of available information.
- **Why?** We map the implementation steps back to the harmonized framework, which represents a synthesis of the recommendations made across 52 reports and the collective expertise of a range of international organizations, government agencies, standards-setting organizations, industry associations, academic institutions, think tanks, and private companies.

Practitioners should use this report as an overarching guide for implementing the wide range of recommended practices necessary to support the responsible adoption of AI systems. The adoption of AI technology affects all aspects of an organization, from its strategy, structure, and operations to its hardware, software, and data. As such, this report provides guidance based on principles from a variety of disciplines. Yet despite their broad impact, AI models and their supporting scaffolding are, at their core, software systems. As a result, many of the recommendations in this guide build on—or draw directly from—existing safety, security, and governance processes that organizations use to manage software systems. While these practices may need to be modified or expanded to address the unique wrinkles that AI models introduce, existing processes and principles should be leveraged by practitioners wherever possible.

Although this report is extensive, it is certainly not comprehensive. The collection of resources concluded in late 2025, and new techniques and methods will continue to be developed. Furthermore, while much of this guidance applies to machine learning (ML) systems broadly, an emphasis is placed on the adoption of generative AI and more advanced frontier models. In addition, the recommendations and resources provided in this report are intended to be use-case and sector agnostic. Organizations should account for the additional nuances specific to their application of AI technology.

We organize this report into four sections. First, we provide some background on the challenges of operationalizing AI best practices and summarize CSET's report on harmonizing AI guidance, which represents the initial step in addressing these challenges. Second, we outline our methodology in developing recommendations and

collating resources. Third, we present the content of the operationalizing guide. Fourth, we close with an assessment of the implications of this work and where future efforts are still needed. Practitioners seeking implementation guidance and resources should feel free to skip directly to the operationalizing guide, which starts on page 18.

Background

The pressure to adopt AI technologies has put many organizations in a challenging position. The purported benefits of AI adoption are high, the need to keep up with competitors is intense, and the risk of AI-enabled threats is growing. The message being received by many organizations is that failing to adopt AI systems risks the organization's reputation, competitive advantage, and security. However, embracing AI technology without the proper policies, procedures, and infrastructure in place to ensure its safety and security presents its own risks that may undermine the intended goals of AI adoption. Despite a myriad of reports, voluntary standards, and proposed best practices for AI systems, there remains a large gap between these recommendations in the abstract and their implementation in practice [11, 12]. The salience of the principle-practice gap has been a common finding across a multitude of studies and reports [1, 2, 13, 14, 15]. The challenge of operationalizing guidance was also a major concern voiced during a workshop hosted by CSET in June 2024 on AI adoption in critical infrastructure [16].

This report represents the second of three phases of research conducted at CSET. This research seeks to reduce the burden of deciphering AI-related guidance currently being shouldered by individual organizations who are seeking to develop or deploy AI technology. The three phases are:

- **Harmonize:** Generate a unified set of recommendations from disparate sources.
- **Operationalize:** Provide steps to implement recommended practices for AI.
- **Tailor:** Apply the guidance to various deployment scenarios and AI use cases.

CSET's recent report entitled "Harmonizing AI Guidance: Distilling Voluntary Standards and Best Practices into a Unified Framework" was the first stage of this work [10]. Now we focus on the second step, operationalizing guidance. In the following sections, we describe how this work relates to the problems organizations face in wrangling existing AI guidance, builds on the previous harmonization report, and identifies future needs in the AI guidance space.

Challenges in Operationalizing AI Guidance

As pressure to adopt AI systems continues to mount, organizations face the daunting challenge of integrating AI technology into their structure, operations, and technical infrastructure. While a variety of voluntary standards and best practices exist to help guide organizations toward this end, implementing proper safety, security, and governance practices remains elusive [1]. As discussed in detail in CSET's recent report

on harmonizing AI guidance, there are five key issues with existing guidance documents that continue to inhibit the successful adoption and integration of AI technology within organizations [10]. First, organizations are inundated with **information overload**. The number of reports and frameworks continues to grow at a prolific rate. Second, these guidance documents come from a wide range of **disparate sources** whose recommendations often overlap and at times conflict with one another. Third, the information contained in these reports is often presented in **inaccessible language** that makes it difficult for organizations to decipher. Fourth, among these documents there are a plethora of high-level objectives identified that organizations are told to achieve, but there is a **lack of implementation details** provided to help guide organizations toward meeting those goals. Fifth, many of these frameworks adopt a **one-size-fits-all** approach, providing a single set of broadly applicable guidance that can frequently be ill-suited for diverse and wide-ranging use cases. For further information on each of these barriers, please refer to the original CSET report on harmonizing AI guidance.

While these challenges stem from the corpus of AI guidance and standards, it is also important to underscore that these issues are compounded by the rapid pace at which large language models (LLMs) and the broader field of artificial intelligence is evolving. As the technology advances, and safety and security practices adapt to keep pace, recommended practices—particularly those that are more granular—become a moving target both for those providing the recommendations and those implementing them. This research does not solve this problem and, as such, acknowledges that this work represents a snapshot of best practices at a specific point in time. That said, AI models are software systems at their core. Many of the established practices and broad principles that have been developed over several decades will continue to apply. These recommendations, like those of other existing guidance documents, will continue to be refined as the technology and state of the art continues to evolve.

CSET Harmonized Framework

To begin addressing these guidance-related barriers, CSET researchers recently developed a harmonized framework for AI development and adoption based on an analysis of over 7,741 recommendations provided across 52 existing guidance documents. This corpus was composed of a variety of frameworks, voluntary standards, and industry best practices that were produced by a range of international bodies, government agencies, standards-setting organizations, industry associations, think tanks, and academic institutions. Using a mix of quantitative and qualitative methods, CSET researchers distilled this collection of knowledge into 258

recommendations, covering 34 topic areas related to governance, safety, security, privacy, and detection and response.

In providing a consolidated, stand-alone, clearly written set of guidance, the harmonized framework helps to address the ongoing challenges related to information overload, disparate sources, and inaccessible language. However, that represents only the first step in removing these guidance-related barriers. This report, representing the second stage of our work, sets out to tackle the fourth identified challenge: the lack of implementation details. This work builds on our previous harmonization efforts, tying concrete actions that an organization can take to the recommendations and objectives contained in CSET framework and, by extension, the broader set of AI guidance that the harmonized framework encapsulates.

Tailoring to Specific Organizational Needs

The recommendations for operationalizing guidance provided in this report are intended to apply broadly to organizations seeking to adopt or deploy AI-based systems for a wide range of use cases. As such, the one-size-fits-all problem remains. The guidance contained in this report will not fully address all of the important aspects of the specific organization, sector, and use case in which the AI system is being deployed. Therefore, use the information provided in this report as a baseline for operationalizing AI safety and security practices. In addition, be sure to assess the nuances of the specific AI use case, identify where this control baseline is insufficient, and seek out further context-specific guidance where available. Some of that guidance may come from future CSET work that plans to tackle various aspects of the one-size-fits-all problem.

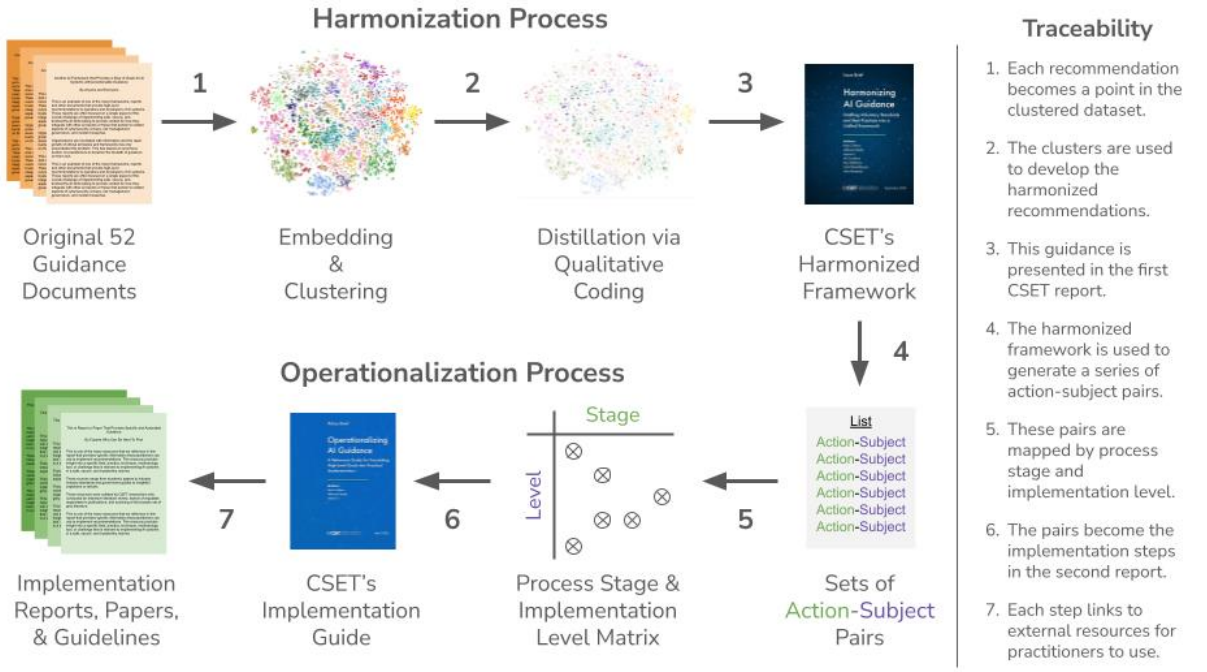
Operationalizing Guidance

The primary aim of this report is to make AI guidance more accessible to organizations by providing a link between the abstract objectives, concepts, and recommendations that make up existing AI guidance and the concrete actions organizations should take to integrate AI systems into their existing policies, practices, and infrastructure. While there are many theoretical approaches for transforming abstract recommendations into concrete practices—often derived from implementation science and applications in the medical field—there is no single agreed-upon process [17]. However, many of these strategies boil down to identifying the fundamental components of the implementation, essentially the questions of who, what, when, where, why, and how [18]. We adopt a generalized version of this approach as follows.

- **What:** We first identify the implementation steps needed for responsible AI adoption by synthesizing the set of *actions* that an organization needs to take and the *subjects* of those actions from CSET's harmonized AI framework [10]. See the Mapping Actions and Subjects section for further details.
- **When:** We then organize these implementation steps by aligning the *actions* with different stages of AI adoption, a process that resembles the software development life cycle (SDLC). See the Stages of AI Adoption section.
- **Where:** In the same vein, we identify the level at which the implementation step occurs based on the *subject*. See the Level of Implementation section.
- **Who:** Next, we assess which functional teams within the organization would be most relevant in implementing these steps. See the Ownership Within the Organization section.
- **How:** For each implementation step, we add more granular recommendations and provide references to external resources for practitioners looking to dive into greater detail.
- **Why:** We provide a link between each implementation step and the recommendations from CSET's harmonized framework. In doing so, we establish a provenance chain from the recommendations in the original 52 reports, to CSET's harmonized framework, to the implementation steps provided in this report, and finally to the breadth of external resources identified alongside our implementation guidance.

We illustrate our end-to-end process of harmonizing and operationalizing guidance in Figure 1. In addition to demonstrating our methodology used across the two CSET reports, this figure shows how the implementation steps and external resources provided in this report trace back to the high-level recommendations from the original 52 guidance documents, as described above.

Figure 1: Overview of the Harmonization and Operationalization Process



Source: CSET.

In the following sections, we describe the process of mapping the harmonized framework into the action-subject pairs used to develop our implementation steps, the stages of AI adoption used to organize the overall structure of this report, the levels of implementation that provide structure within each process stage, and finally an overview of the different functions within the organization and their general responsibilities for AI adoption. Alternatively, you can jump directly to CSET's Guide to Operationalizing AI Systems.

Mapping Actions and Subjects

To identify the key implementation steps that organizations should undertake in their adoption of AI technology, we analyzed the recommendations from the harmonized framework and broke them down into a series of action-subject pairs. To illustrate how this works, take the example of recommendation IT.1 from the harmonized framework, which reads: "Articulate the organization's mission and develop a strategy to achieve its objectives. Align [information technology] IT initiatives and systems with organizational goals."

We break this recommendation down into four action-subject pairs. The first two can be derived from the initial sentence, becoming "articulate mission" and "develop

strategy." We segment the second sentence in much the same way, except that the action is the same for both pairs in this case, leaving us with "align IT initiatives" and "align systems." After applying this methodology to each recommendation, we end up with 814 action-subject pairs. Although that is quite extensive, many of these pairs are closely related or overlap. Through a process of qualitatively grouping similar actions and subjects, we narrow down this set to 232 distinct pairs, each of which corresponds to one of the implementation steps in this report.

Stages of AI Adoption

In analyzing the action-subject pairs, it quickly became apparent that the set of *actions* roughly aligned with the typical stages of the software development life cycle (SDLC). Therefore, we employ a modified version of this process, which we call the AI adoption life cycle. Figure 2 illustrates the process stages and substages that emerged from the analysis.

Figure 2: The Stages of AI Adoption

Source: CSET.

We structure the guidance provided in this report according to the stages and substages of the life cycle diagram. However, it is important to note that while the report is organized sequentially to convey the information in a clear manner, some steps of the process represent ongoing activities, such as operations and monitoring,

and others should likely occur at multiple points throughout the process, such as approval and communication. The stages and substages are described in greater detail below.

- **Review:** The review stage serves as the entry point for the organization in the AI adoption cycle. To begin, the organization should *self-assess* its maturity and readiness to adopt AI technology. At the end of the cycle, the organization should review its policies and procedures to identify areas where it can *improve*. The organization should implement those improvements before continuing the cycle by conducting another self-assessment.
- **Planning:** After conducting a self-assessment, the organization moves into the planning phase. During this stage, the organization should *define* its objectives and requirements for the AI system being adopted, *plan* how the organization will acquire and support the technology, *promote* norms and standards in organizational and system design, *map* how the system will integrate into the organization, and *assess* the potential risks and impacts of adopting the AI system.
- **Implementation:** In the first two steps of the implementation stage, the organization will *acquire* or *develop* the AI system. The extent to which the organization relies on acquisition versus development can vary widely. On one end of the spectrum, an organization could opt to acquire a fully-fledged AI product or service directly from a third party. On the other end, an organization could collect data, train a model, and develop a system using that model all in-house. Many use cases are likely to fall somewhere in the middle, with organizations acquiring some components and developing others themselves. Regardless of where the organization lands on this spectrum, it will need to *control* the access to—and the behavior of—the system and *protect* the system from external threats. Finally, the organization must *test* the performance, capabilities, and vulnerabilities of the system before making a decision whether to *approve* its deployment.
- **Deployment:** During this stage, the organization must first *communicate* relevant information about the system and its development, both internally and externally. Then the organization will *deploy* the AI system, integrating it into the existing infrastructure and operating procedures.
- **Operation:** The operation stage represents the ongoing activities to maintain the proper functioning and performance of the AI system. During this stage the organization should *engage* relevant stakeholders, *manage* the performance of the system over its lifespan, *monitor* for potential threats, and *respond* to issues when they arise.

Level of Implementation

In a similar manner to how the *actions* in the action-subject pairs neatly mapped onto the process stages of the SDLC, the *subjects* presented an inherent structure that corresponded to various implementation levels within the organization. These implementation levels fall roughly under the jurisdiction of various leadership roles within the organization, providing some insight into the actors responsible for operationalization. We describe each of these levels below.

- **Organization:** The organization level consists of the physical and external environment in which the organization operates, the organization's internal structures, and the high-level strategy that governs the adoption of AI technology within the organization. As the broadest level of implementation, these actions generally fall within the purview of the chief executive officer.
- **Operations:** The operations level represents the day-to-day activities of the organization. At this level, implementation focuses on the development of operating procedures and contingency plans, the operation of various organization-wide functions (e.g., legal, compliance), and the fostering of relationships with external stakeholders. Many, but not all, of these responsibilities tend to align with that of the chief operating officer.
- **Workforce:** The workforce level revolves around the organization's personnel. Guidance at this level focuses on fostering and hiring the requisite talent for AI adoption, as well as overseeing the activity of personnel within the organization. In large part, this level corresponds to the duties of the chief human resources officer.
- **Infrastructure:** The infrastructure level consists of the architecture, network, and services that form the information technology backbone of the organization. This level incorporates the broad technical initiatives that will support the organization's adoption and integration of AI systems. Responsibilities at this level generally fall under the chief information officer.
- **Data:** The data level focuses on the collection, processing, and management of data—particularly in the context of data being used to train and fine-tune AI models. We consider this to be a level above the AI model and system, as a single dataset may be used to train multiple models. Depending on the structure of the organization, these activities may be within the scope of the chief data officer or may be rolled up into that of the chief information officer.
- **Model:** The model level consists of the set of activities surrounding the training and testing of the AI model itself. As such, implementation at this level is likely to be the most novel for organizations. Like before, we opt to order the model

above the system, as a single model might be deployed as a part of multiple systems. Ownership at the model level may fall across various roles in the organization. Given the focus on innovation and development, much of the responsibility may fall to the chief technology officer. However, the chief AI officer may take on these duties if an organization decides to establish a new board-level position specifically to oversee the use AI technology within the organization.

- **System:** The system level represents the implementation of the end product—which, for the purposes of this report, is built around an AI model—that users within the organization or customers of the organization will use. As such, implementation at this level focuses around the software scaffolding built around the model and how end users will interact with the system. Responsibility at the system level is likely to vary across organizations, but in many cases will fall to a specific product team under the oversight of a chief product officer.

Ownership Within the Organization

While ownership over implementation can be ascribed broadly by level, in reality responsibility for implementing these practices will not be nearly as neatly defined. Many teams and functional groups throughout the organization will be involved in activities across a range of levels and stages. In addition, we fully acknowledge that organizations, and their org charts, vary widely. As such, we hesitate to be too prescriptive about identifying ownership. However, ignoring the roles and responsibilities for implementing guidance—which many frameworks opt to do—does a disservice to organizations looking for more granular details. Therefore, alongside the implementation guidance provided in this report, we include an indicator of the functional teams (e.g., legal, cybersecurity, research and development) within the organization that are likely to be involved in operationalizing it. We do not attempt to provide more granular distinctions at the level of individual positions or specific teams. We summarize the relevant functional teams and their responsibilities in Table 1 below.

Table 1: Functional Teams and Their Role in Operationalizing AI Guidance

Team	Leadership	Role
Leadership	Chief Executive Officer (CEO)	Defining objectives for AI adoption and making final approval decisions
Strategy	Chief Strategy Officer (CSO)	Establishing a strategy, related policies, and high-level requirements for AI initiatives
Finance	Chief Financial Officer (CFO)	Prioritizing and allocating adequate resources to support AI-related activities
Risk	Chief Risk Officer (CRO)	Assessing and managing the overall risk introduced by the adoption of AI systems
Operations	Chief Operations Officer (COO)	Developing operating procedures for AI systems and overseeing the implementation of AI policies
Workforce	Chief Human Resources Officer (CHRO)	Hiring and retaining requisite talent to responsibly adopt AI technologies
Compliance	Chief Compliance Officer (CCO)	Ensuring the organization meets its AI-related obligations as intended
Responsibility	Chief Responsibility Officer (CRO)	Promoting ethical and responsible AI practices in collaboration with partners
Physical Security	Chief Security Officer (CSO)	Protecting and monitoring the facilities and physical environment of the organization
Communication	Chief Communications Officer (CCO)	Communicating AI practices, decisions, and outcomes internally and externally
Marketing	Chief Marketing Officer (CMO)	Advertising AI products and services transparently and engaging with end users
Supply Chain	Chief Supply Chain Officer (CSCO)	Vetting and coordinating suppliers of AI-related data, models, tools, and services
Cybersecurity	Chief Information Security Officer (CISO)	Protecting AI systems, managing access, and monitoring of security-related events
Privacy	Data Privacy Officer (DPO)	Controlling the use of personal or sensitive data in model training and operation
Trust & Safety	Chief Trust & Safety Officer (CTSO)	Preventing harmful output and behavior of the AI system and assessing AI impacts
Infrastructure	Chief Information Officer	Managing the infrastructure services that support the

Team	Leadership	Role
	(CIO)	operation of the AI system
Data	Chief Data Officer (CDO)	Managing data used to train models and evaluating issues of quality and bias
Innovation	Chief Technology Officer (CTO)	Overseeing the research and development of new AI models and systems
Product	Chief Product Officer (CPO)	Implementing and managing the AI system, from ideation to decommissioning

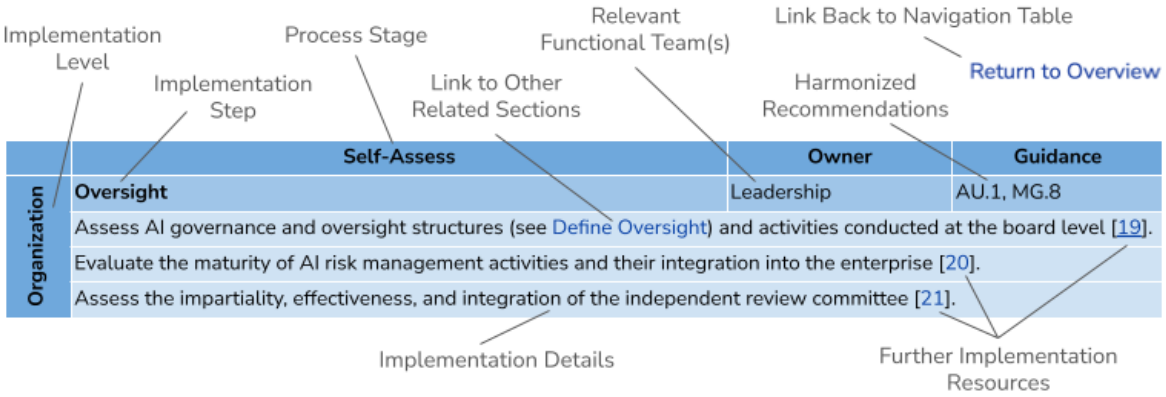
CSET's Guide to Operationalizing AI Systems

In this section, we present an extensive reference guide to aid organizations in operationalizing AI best practices. The guide is based on an extensive review of government frameworks, industry standards and best practices, academic papers, and grey literature sources by CSET researchers to identify practical content that can help organizations translate high-level guidance into implementation steps. We organize our guide into sections that correspond to each stage in the AI adoption process. This starts with a self-assessment of the organization and its readiness to adopt AI technology; proceeds through the planning, development, and operation of the AI system; and concludes with the organization's improvement of its own practices and capabilities. Within each section, we provide the recommendations in a tabular format organized by the level at which they should be implemented, starting with organization as a whole and progressing down to the AI system level.

Since the process of adopting AI technology touches on almost all major aspects of the organization, one of the key challenges in developing this guide was balancing the breadth and depth of the information provided. To strike the right balance, we highlight the most important information related to each implementation step within this report and provide references to more detailed resources for practitioners to dive deeper into specific areas. As such, the citations included alongside each implementation step differ from the references in other reports. Instead of providing evidence to support why we provide a given recommendation, the citations supply direct links to external resources that elaborate further on a specific topic, practice, technique, or tool that can be used to implement the given recommendation.

In addition to these external resources, we identify the relevant functional teams within the organization that will likely be involved and provide references to the recommendations from the harmonized framework that the step helps to implement. In Figure 3, we provide an overview of how this information is presented throughout the guide.

Figure 3: Layout of the Each Implementation Step in This Guide



Source: CSET.

How to Navigate This Guide

The content of this guide is extensive, covering adoption practices across an organization. For many practitioners, only a subset of this information may be applicable. To aid in the identification of relevant content, we provide a complete overview of the guide in Table 2, along with links for quick navigation. Below, we lay out several methods of searching for pertinent information and provide examples of when they might be useful to different types of practitioners.

By process stage: Navigate directly using the links in the first column of Table 2.

- A board member mapping out a vision for the organization's AI adoption
- A CAIO building out the organization's AI capabilities from scratch

By implementation level: Use the links within Table 2 under the relevant column.

- An HR manager looking for guidance on AI workforce development and hiring
- A privacy professional concerned about data practices related to an AI system
- A project lead ensuring their AI system incorporates proper safeguards

By functional team: Search the document using the names defined in Table 1.

- A legal professional assessing the legal ramifications of adopting an AI system
- A risk management specialist identifying and measuring AI-related risks
- A financial officer determining the allocation of resources for AI initiatives

By recommendation: Search by ID used in the harmonized framework (e.g., IT.1, SC.3).

- A compliance officer ensuring practices conform to existing standards

Table 2: Overview of the Topics Covered in CSET's Guide to Operationalizing AI (with links for navigation)

	Organization	Operations	Workforce	Infrastructure	Data	Model	System
Self-Assess	• Strategy • Oversight • Culture	• Practices • Relationships • Resources	• Talent	• Architecture • Innovation • Security & Privacy	• Data Management	• Performance • Test & Evaluation	• Impact & Trust
Define	• Context • Requirements • Objectives • Use Cases • Oversight	• Essential Function • Incidents • Constraints	• Needs	• Classification Scheme	• Legal Basis	• Release Criteria	• Purpose • System Requirements • Metrics
Plan	• Strategy • Policies	• Resource Allocation • Life Cycle • Response	• Staffing & Support	• Integration	• Storage & Retention	• Testing • Release	• Design • Deployment
Promote	• Culture • Responsibility	• Standard Practices • Compliance	• Accountability	• Common Architecture • Least Privilege & Functionality	• Data Minimization	• Safety • Trust • Human Oversight	• Privacy • Security • Ethical & Safe Use
Map	• Stakeholders	• Supply Chain	• Ownership • Approval	• Assets • Data Flows	• Classification • Provenance	• Traceability	• Components
Assess	• Points of Failure • Threats	• Risk • Legality • Ethics • Supply Chain	• Capability	• Platforms	• Data Protection	• Technology Readiness • Trade-Offs	• Impact • Attack Surface
Acquire	• Relationships	• Due Diligence • Contracts	• External Expertise	• Verification	• Consent • Data Collection • Third-Party Datasets	• Third-Party Models	• Third-Party Systems

	Organization	Operations	Workforce	Infrastructure	Data	Model	System
Develop	• Responsible Innovation	• Management Systems	• Diverse Teams	• Logging • Version Control	• Data Quality	• Alignment • Robustness • Transparency • Federated Learning • Privacy Preservation	• Secure Software • Patches
Control	• Physical Environment	• Risk	• Activity	• Identities • Access • Data Flows	• Data Transfers • Classified & Proprietary Data	• Bias • Inputs • Outputs • Synthetic Media	• Action Space • Agency
Protect	• Physical Premises	• Intellectual Property	• Personnel	• Network • Assets • Logs • Backups	• Confidentiality • Integrity	• Weights • Inputs & Outputs • Execution • Resource Utilization	• Availability • Source Code • Users
Test	• Business Continuity	• Incident Response	• Competency	• Controls • Updates • Restoration	• Bias & Skew	• Capabilities • Performance • Fairness	• Security • Safety • Quality • Acceptance • User Interaction
Approve	• Go/No-Go Decision	• Stakeholder Input • Risk Threshold	• Independent Validation	• Integration & Acceptance	• Privacy Assurance	• Release Criteria	• System Review

	Organization	Operations	Workforce	Infrastructure	Data	Model	System
Communicate	• Objectives • Policies	• Practices • Decisions	• Responsibilities	• Controls • Changes	• Data Practices • Data Transparency	• Risks & Limitations	• Test Results
Deploy	• Review Cycle	• Enforcement	• Training & Drills	• Architecture • Inventory & Mappings	• Disclosure	• Release	• Deployment
Engage	• Government • Society • Academia	• Stakeholders	• Personnel	• Service Providers	• Data Subjects	• Collaborative Initiatives	• End Users
Manage	• Situational Awareness	• Relationships	• Personnel Awareness	• Assets • Access • Maintenance • Decommission	• User Control	• Performance	• Vulnerabilities
Monitor	• Context • Physical Environment	• Centralized Analysis • Information Sharing • Third Parties	• Security Operations • Insider Threats	• Access • Data Flows	• Data Quality • Privacy	• Performance • Alignment • Fairness • Inputs & Outputs	• Behavior • Use • Impact
Respond	• Remediation	• Triage • Communication	• Response Team • Alerting	• Containment & Neutralization • Investigation	• Data Breach • Restoration	• Fail-Safes • Restoration	• Resilience Mechanisms • Restoration
Improve	• Strategy	• Procedures • Monitoring • Response • Supply Chain	• Training & Awareness	• Controls	• Data Handling	• Outcomes • Testing	• System Design • User Experience

Self-Assess

The Self-Assess stage represents the entry point into the operational life cycle. Here, the organization should assess its readiness to adopt AI technology [19]. We recommend using an AI maturity model—such as those provided by the General Services Administration, CNA, OWASP, MITRE, and others—to conduct a broad self-assessment [20, 21, 22, 23]. There are also several publicly available self-assessment tools, including those from the U.S. Chief Digital and Artificial Intelligence Office, the Commission Nationale de l'Informatique et des Libertés, Cisco, the European Institute of Innovation and Technology, and Fifth Quadrant [24, 25, 26, 27, 28]. Less mature organizations should start with simpler ML applications and leverage third-party options before attempting to deploy or build more advanced models.

	Self-Assess	Owner	Guidance
Organization	**Strategy**	Strategy	SL.1
	Assess how well the AI strategy (see Plan Strategy) aligns with organizational goals and priorities [29].		
	Evaluate the strategic alignment of organizational initiatives using established models [30, 31, 32, 33].		
	Identify the frequency and regularity with which the AI strategy is reviewed and updated [34].		
	Oversight	Leadership	AU.1, MG.8
	Assess AI governance and oversight structures (see Define Oversight) and activities conducted at the board level [35].		
	Evaluate the maturity of AI risk management activities and their integration into the enterprise [36].		
	Assess the impartiality, effectiveness, and integration of the independent review committee [37].		
	Culture	Leadership, Workforce	SL.5, WF.5, AU.6
	Determine the organization's cultural readiness (see Promote Culture) to adopt new technologies [38].		
	Conduct employee surveys to measure the degree to which AI norms have been embraced [39].		
Operations	**Practices**	Operations	MG.3
	Evaluate the maturity of business process management activities [40]. Assess the organization's AI capabilities and needs [41].		
	Identify whether the organization maintains operating procedures (see Promote Standard Practices) for the use of AI technology [42].		
	Measure internal compliance with organizational policies and procedures [43].		

	Self-Assess	Owner	Guidance
Operations	**Relationships**	Communication, Supply Chain, Leadership	ST.1, SC.1
	Assess the maturity of stakeholder relationship management (see Manage Relationships) within the organization [44].		
	Evaluate the maturity of the organization's supply chain management [45, 46].		
	Assess participation in collaborative AI initiatives such as AI incident reporting, voluntary commitments, and industry forums [47, 48].		
	Resources	Finance	IT.5, IM.6
	Determine whether there are sufficient financial resources for organizational change to the AI system throughout its life cycle [49].		
	Assess the maturity of enterprise resource management activities and systems within the organization [50, 51].		
Workforce	**Talent**	Workforce	WF.1
	Identify existing talent and expertise within the organization that can help to support the adoption of AI technology [52].		
	Assess the AI competencies of existing personnel (see Assess Capability) and identify workforce gaps (see Define Needs).		
	Determine whether the organization's talent acquisition and training resources can adequately address AI talent shortages [53].		
Infrastructure	**Architecture**	Infrastructure	IT.4
	Assess the maturity of the organization's enterprise architecture [54, 55]. Determine how well it can support new AI technologies [56].		
	Determine how well enterprise architecture aligns with and supports business goals, including those specifically for AI adoption [57].		
	Evaluate the maturity of the organization's adoption of zero trust architecture [58].		
	Identify the availability of AI platforms within the organization and potential cloud platforms that could support AI adoption [59].		
	Innovation	Innovation	IT.6, SI.3
	Evaluate the maturity of the software development process and life cycle (see Plan Life Cycle), also known as DevOps [60, 61].		
	Examine the application of software development processes for AI software development, also known as MLOps [62, 63, 64, 65].		
	Determine the maturity of project management activities within the organization [66].		

	Self-Assess	Owner	Guidance
Infrastructure	**Security & Privacy**	Cybersecurity, Privacy	SM.1, PP.2
	Evaluate the maturity of cybersecurity practices within the organization [67, 68].		
	Determine the maturity of the organization's software security and assurance practices [69].		
	Assess the organization's cybersecurity logging and monitoring capabilities [70].		
	Measure the capabilities and maturity of privacy practices within the organization [71, 72, 73].		
Data	**Data Management**	Data	SM.8, TR.5
	Determine the maturity of data governance activities within the organization [74].		
	Assess the capabilities of the organization in identifying and managing bias within datasets and AI models [75].		
Model	**Performance**	Product	PM.2
	Assess quality and performance of AI models used, or being considered for use, within the organization [76].		
	Evaluate the integration of model performance monitoring and alerting into AI operations and oversight [77].		
	Test & Evaluation	Innovation	TE.1
	Determine the maturity of the organization's software testing and evaluation capabilities [78].		
	Evaluate the test and evaluation capabilities for AI models [79].		
	Identify quality assurance processes and evaluate their effectiveness [80].		
System	**Impact & Trust**	Trust & Safety	IM.1, ST.6
	Determine the maturity of AI ethics and related functions within the organization [81, 82, 83].		
	Evaluate the organizational processes for conducting impact assessments of AI systems [84].		
	Assess the procedures within the organization and characteristics of AI systems that support trustworthiness [85].		
	Evaluate the technical methods, analysis, proxy mechanisms, and processes used to establish explainability for AI systems [86].		

Define

The Define stage occurs at the start of the organization's planning for AI adoption. At this stage, the organization should determine its objectives for AI adoption, the requirements of the AI system, and the constraints within the organization that may affect the safe and secure adoption of AI systems. While most of these responsibilities fall to the organization's leadership, input from all levels is needed to adequately inform the definition of the organization's vision for AI adoption.

	Define	Owner	Guidance
Organization	**Context**	Strategy, Legal	IT.2, AU.2, IM.1, RM.1, PP.1, IR.1
	Map out the context in which the organization operates [87, 88]. Assess the external factors relevant to AI adoption [89].		
	Consider using SWOT [90] and PESTLE [91] analysis to assess the internal and external context of the organization.		
	Identify AI regulation and legislation for jurisdictions in which the organization operates [92, 93]. Also identify relevant privacy laws [94].		
	Define the organization's internal and external stakeholders (see Map Stakeholders).		
	Requirements	Leadership, Legal	MG.1, RM.2, ST.1, RM.2
	Conduct regulatory mapping or modeling to link legal obligations to organizational requirements [95, 96].		
	Gather input from internal and external stakeholders to inform the generation of organizational requirements [97].		
	Define the organization's overarching risk appetite and tolerance, as well as that for individual AI systems [98, 99].		
	Objectives	Leadership	IT.1, MG.2, RM.2, RR.1
	Articulate a mission and vision statement for the organization [100, 101].		
	Incorporate stakeholder and business requirements (see Define Requirements) into the strategic goal-setting process [102].		
	Define the objectives and goals of the organization using tools such as balanced scorecards [103] or the SMART framework [104].		
	Determine the organization's AI objectives and align them with organizational goals [105].		

	Define	Owner	Guidance
Organization	**Use Cases**	Strategy, Innovation	IT.1, IT.5, IM.2
	Identify existing challenges and new opportunities within the organization where AI adoption may be impactful [106, 107, 108].		
	Categorize and assess the AI use case [109]. Determine if AI is appropriate for the given challenge or opportunity identified [110].		
	Work with stakeholders and potential end users to validate use cases early on in the development life cycle [111].		
	Oversight	Leadership	SL.2, AU.1, RC.1, MG.8
	Define the board's responsibilities for AI governance [112]. Ensure AI oversight [113] and AI knowledge [114] at the board level.		
	Develop the appropriate level of AI [115] and cybersecurity [116] knowledge among the organization's leadership.		
	Designate ownership of AI among the organization's top-level executives and identify champions throughout the organization [117].		
	Consider creating a new CAIO board-level position [118, 119, 120].		
	Establish independent oversight over AI use within the organization. Consider formalizing an independent review committee [37].		
Operations	**Essential Function**	Leadership, Operations	SM.3, RR.1
	Define the critical business functions within the organization, including any government-defined essential functions [121].		
	Identify the dependencies on which the essential function relies (see Map Supply Chain and Map Assets).		
	Incidents	Cybersecurity, Privacy, Trust & Safety, Operations	IR.3
	Assess the range of potential incidents, including cyber [122, 123], privacy [124], AI [125, 126, 127], and operational [128] incidents.		
	Specify a set of criteria for each incident that, if met, would trigger a planned incident response within the organization [129, 130].		
	Develop an incident priority matrix that maps potential incidents across severity (impact) and priority (urgency) [131].		
	Constraints	Finance, Infrastructure	IT.5, RR.4
	Identify the organization's resource constraints (e.g., physical, human, financial, informational, time) [132].		
	Evaluate the organization's AI resources and constraints related to the AI triad [133] (data, algorithms, compute) and workforce [134].		

	Define	Owner	Guidance
Workforce	**Needs**	Workforce	IM.3, ST.4
	Identify the skills and competencies required to support the organization's adoption of AI technology [135, 136, 137].		
	Determine the AI talent needs of the organization [138, 139] and compare against existing internal expertise (see Self-Assess Talent).		
	Consider the leadership and talent needed to promote a positive culture for AI governance [140].		
	Understand the AI talent market, including the supply [141, 142] demand [143], and gap [144] for AI-related talent.		
Infrastructure	**Classification Scheme**	Infrastructure	IS.1
	Define a data and asset classification scheme that captures relevant security, privacy, and handling information [145].		
	Extend the classification scheme to the AI system, incorporating metrics from existing classification tools and efforts [146, 147].		
	Document the classification scheme and establish a policy that guides its implementation across the organization [148].		
Data	**Legal Basis**	Legal, Privacy	PI.2
	Assess how personal data is being processed within the organization (see Map Data Flows).		
	Identify the lawful basis for each processing activity involving personal data [149, 150].		
	Consider the additional complexity of defining a lawful basis for processing when AI technology is involved [151].		
Model	**Release Criteria**	Innovation	IT.7, TO.2
	Set release criteria that the model and system must meet prior to receiving approval for deployment [152].		
	Identify factors influencing the decision to deploy, or to even pursue developing, a given technology [153].		
	Set risk thresholds for the performance and capabilities of AI models to govern release and response decisions [154, 155, 156].		
System	**Purpose**	Innovation	IT.7, PI.2
	Articulate the specific challenge that the proposed system is intended to address or the benefit that it is intended to provide [157].		
	Define the goals of the system and establish ways to track whether the system is meeting those goals [158].		

	Define	Owner	Guidance
System	**System Requirements**	Innovation, Product	IT.7, IT.8, MG.1, TO.2, IS.1
	Use requirements engineering principles to elicit system requirements from a diverse set of relevant stakeholders [159, 160, 161].		
	Consider the factors and challenges in conducting requirements elicitation for AI systems [162, 163].		
	Identify and resolve tensions that arise between (frequently technical and socio-technical) requirements [164, 165].		
	Metrics	Innovation, Product	PM.1, FS.2, SG.4, TE.6
	Identify metrics to measure the ability of the system to meet requirements (see Define Requirements and Define System Requirements).		
	Define metrics and key performance indicators to track the AI model's performance over time [166, 167, 168].		
	Include measures that capture a wide range of goals such as fairness [169], explainability [170], privacy [171], and sustainability [172].		
	Align performance metrics with the needs of users and the organization's objectives (see Define Objectives) [173].		
	Understand the benefits and limitations of existing AI benchmarks [174, 175, 176].		

Plan

The Plan stage involves the creation of a strategy, policies, and plans to implement the organization's vision for AI adoption as defined in the previous stage. As indicated by the breadth of this report, AI adoption affects almost all aspects of the organization in some way, shape, or form. Therefore, practitioners must not only develop plans for the AI system specifically but also update other plans to account for AI adoption and use within the organization.

	Plan	Owner	Guidance
Organization	**Strategy**	Strategy, Leadership	IT.1, MG.2, RM.2, SL.1, TE.1, PI.1, MO.2, RR.1
	Define a business strategy that articulates the organization's goals and priorities for the development and use of AI technology [177].		
	Align the AI strategy with the organization's overarching mission and goals (see Define Objectives) [29].		
	Ensure that responsible AI initiatives are prioritized as a part of the organization's strategy [178].		
	Policies	Leadership	AU.5, MG.4, MG.6, RC.1, RM.7, SC.5, SI.1, WF.4, AC.1, MS.5, PI.9
	Create an organization-wide policy for AI that guides AI development and deployment within the organization [179, 180, 181].		
	Include a policy on acceptable use of AI within the organization to guide personnel on what activities are allowed and prohibited [182].		
	Establish and enforce an acceptable use policy for internal and external use of the organization's AI tools and services [183].		
	Integrate relevant AI guidelines into other organizational policies such as that for cybersecurity, privacy, data, and risk management [184].		
Operations	**Resource Allocation**	Finance	IT.5, IR.2, WF.3, MG.3, LG.6, DD.1, IM.6
	Plan the allocation of financial, physical, technological, and human resources in alignment with strategic priorities [185].		
	Account for the unique aspects of AI systems, including their high computational intensity and related environmental impact [186].		

	Plan	Owner	Guidance
Operations	**Life Cycle**	Innovation	IT.6
	Establish life cycle stages (similar to the process stages in this report) that account for AI [187, 188] and software [189] development.		
	Build approval processes into the development life cycle, defining where organizational and independent oversight is required [190].		
	Apply software engineering approaches to the development and operation of AI systems [191].		
	Identify AI governance tools across life cycle stages. Note that gaps exist outside of design and development [192].		
	Response	Operations	RC.3, VN.6, ES.5, IR.1, SC.2, RR.2, RR.7, RM.7
	Institute an organization-wide incident response plan that accounts for cybersecurity, privacy, and safety incidents [193, 194, 195].		
	Develop a specific plan for responding to incidents that involve or directly stem from an AI system [196].		
	Conduct a business impact analysis [197] to inform the development of a resilience plan [198] and an IT disaster recovery plan [199].		
	Develop a risk mitigation plan based on the identification and assessment of risks posed by the AI system (see Assess Risk) [200].		
Workforce	**Staffing & Support**	Workforce	DD.1, WF.5, ES.4, SI.3
	Develop a skills-based hiring strategy for acquiring the AI-related talent that the organization needs [201].		
	Conduct workforce planning [202], accounting for the range of technical [203] and soft skills [204, 205] needed to support AI adoption.		
	Plan for the potential displacement of personnel [206] due to the adoption of AI technology [207, 208, 209] in an ethical manner [210].		
	Develop a reskilling program to integrate personnel into the AI-enabled workforce [211, 212], accounting for program limitations [213].		
Infrastructure	**Integration**	Infrastructure	IT.4, SL.4, PP.2, LG.3
	Assess how the AI system aligns with the organization's enterprise architecture (see Promote Common Architecture).		
	Create a plan for how the AI system will integrate into the organization's existing infrastructure [214].		
	Plan how logging and monitoring of the AI system will integrate into the broader organization-wide activities (see Deploy Architecture).		
Data	**Storage & Retention**	Data	IS.1, MG.6, LG.6
	Plan how data used in model training and generated during operation, logging, and monitoring will be stored in a secure manner [215].		
	Determine the capacity requirements for the system's data needs and identify ways of adjusting that capacity as necessary [216].		
	Establish a data retention policy that specifies how long personal data will be stored and how it will be securely disposed of [217, 218].		

	Plan	Owner	Guidance
Model	**Testing**	Innovation	TE.1
	Establish an organization-wide testing strategy for AI systems and their performance [219, 220].		
	Take a multipronged approach to testing that incorporates strategies and paradigms from a variety of disciplines [221].		
	Include AI red teaming in test and evaluation activities [222]. Leverage available red-teaming techniques and tools [223].		
	Release	Innovation	IT.8, IS.9
	Document a set of criteria that the model must meet before being released (see Define Release Criteria).		
	Plan how the model will be released and who will have access [224], taking into account external versus internal deployment [225].		
	If released publicly, determine the degree of model openness and assess the trade-offs in benefits and risks [226, 227, 228, 229].		
System	**Design**	Innovation, Product	IT.7
	Scope the design, and degree of in-house development [230], based on available resources (see Plan Resource Allocation).		
	Develop a plan for developing and managing the system's security, privacy, and supply chain throughout its life cycle [231].		
	Understand the limitations and unresolved challenges in technical AI governance when designing the system [232].		
	Follow software [233] and ML [234] design principles, as well as emerging best practices for AI design [235, 236].		
	Consider the range of additional principles that should be embedded into the system's design (see Promote).		
	Deployment	Innovation, Product	DD.8, IT.8, IT.9
	Document the set of criteria that the system must meet before being deployed (see Define Release Criteria).		
	Develop a strategy for the system's deployment that uses a staged approach, such as blue-green or canary deployments [237].		
	Determine the scope of systems and users that the AI system will be rolled out to at each stage of the deployment [238].		

Promote

The Promote stage serves a twofold purpose. First, it involves the promotion of positive values and practices within the organization. Second, at a more technical level, it includes the promotion of design principles within the infrastructure and systems employed by the organization. As such, at the lower implementation levels this stage corresponds to the design phase of many software development life cycles.

	Promote	Owner	Guidance
Organization	**Culture**	Leadership, Workforce	AU.6, MG.2, SL.5, WF.5
	Take actions that facilitate the development of a positive, risk-aware culture within the organization [239].		
	Develop an organizational culture that promotes the AI adoption [240, 241] and its safe and responsible use [242].		
	Assess factors influencing the organization's cybersecurity culture [243, 244] and develop structures that foster a positive one [245].		
	Build an organizational privacy culture and climate [246, 247]. Identify and support privacy champions within the organization [248].		
	Responsibility	Responsibility, Legal	RC.3 SI.1, SI.3, SI.5, SI.6
	Establish an ethics program within the organization [249, 250]. Consider creating an AI ethics board for independent oversight [251].		
	Understand the challenges that ethics owners within the organization face and work to empower and support them [252].		
	Ensure that the design of AI systems conforms with ethical AI principles [253].		
	Foster an ethical culture [254] and implement due diligence practices to ensure responsible business practices [255].		
	Create confidential mechanisms to report ethical, legal, and safety concerns or violations [256].		
	Account for the environmental impact of AI technology and its relation to the organization's sustainability goals [257, 258].		
Operations	**Standard Practices**	Operations	IT.1, IT.6, MG.1, MG.4, SC.1, LG.1, DD.1, SL.4
	Align practices with strategy (see Plan Strategy) and policies (see Plan Policies).		
	Develop standard practices for AI development and operations that align activities across business functions [259].		
	Apply operations management principles in implementing processes to govern and manage AI activities within the organization [260].		
	Identify key AI-related tasks and develop operating procedures to standardize how they are carried out [261].		

	Promote	Owner	Guidance
Operations	**Compliance**	Compliance	AU.2, AU.4, SC.4, SI.7, MS.8, PI.10, PM.5
	Develop a program to promote compliance with organizational policies (see Plan Policies) and regulation (see Assess Legality) [262].		
	Customize policies to fit the needs of the organization and design them to be realistic and usable by employees [263].		
	Use employee performance feedback as a mechanism to promote compliance with organizational AI policies [264].		
	Work with personnel and business partners to identify and address reasons for noncompliance [265].		
Workforce	**Accountability**	Compliance, Workforce	RC.4, SL.2, WF.4, TO.7
	Collaborate with teams and individual personnel in defining responsibilities (see Map Ownership) and expectations [266].		
	Develop an accountability framework to empower, guide, and track the success of decentralized teams [267, 268].		
	Foster a culture of support and accountability within teams and between managers and direct reports [269, 270, 271].		
Infrastructure	**Common Architecture**	Infrastructure	IT.4, DD.9, DD.10
	Develop a common enterprise architecture across the organization [272], accounting for AI-specific needs [273].		
	Establish a zero trust architecture that verifies all users, devices, and systems when attempting to access resources [274, 275].		
	Compile a common set of controls that systems within the organization should implement [276].		
	Adhere to a common set of security configurations for organizational assets and applications [277].		
	Least Privilege & Functionality	Cybersecurity	AC.2, AC.7, IS.8
	Adhere to the principles of least privilege [278] and least functionality [279] in implementing access control.		
	Apply these principles when specifying access to a given system by personnel or third parties [280].		
	Apply these principles when specifying what a given system can access, which is particularly important for agentic AI [281].		
Data	**Data Minimization**	Data, Privacy	PI.2, PP.6
	Minimize the collection, processing, retention, and transfer of personal or personally identifiable information (PII) [282].		
	Extend the principle of data minimization to the collection of data used for training and fine-tuning AI models [283, 284].		
	Consider trade-offs between minimization and the collection of data to combat bias and monitor use (see Assess Trade-Offs).		

	Promote	Owner	Guidance
Model	**Safety**	Trust & Safety	SI.4, SI.5
	Conduct an early analysis of potential mismatches between needed and actual AI performance that may lead to harm [285].		
	Incorporate safety guardrails into the AI system at the design phase [286].		
	Consider employing a human-centered design approach for AI systems [287].		
	Trust	Trust & Safety	SI.2, IM.3, IM.7
	Establish a meaningful participatory approach to AI design that incorporates stakeholder input and works to build trust [288].		
	Integrate ethical [289] and trustworthy [290] principles into the design of AI systems.		
	Consider employing a value sensitive design approach, adapted for AI, in the requirements gathering and design phase [291].		
	Human Oversight	Trust & Safety	TO.3
	Carefully consider the role that a human will play in the operation of the system, particularly when critical decision-making occurs [292].		
	Design oversight to be human-centric, emphasizing collaboration (AI-in-the-loop) rather than automation (human-in-the-loop) [293].		
	Design AI systems that can identify, measure, and communicate the level of competency or uncertainty in their output [294, 295].		
	Explore the range of human-machine teaming concepts that should be considered as a part of the system's design [296, 297].		
System	**Privacy**	Privacy	DD.1, PP.4, PI.1
	Promote a privacy-by-design approach to system development within the organization [298].		
	Apply established privacy principles to the design of AI systems [299, 300].		
	Security	Cybersecurity	DD.1, DD.3, PP.4, SI.4
	Promote a security-by-design approach to system development within the organization [301].		
	Apply established security principles to the design of AI systems [302], particularly LLM-based [303] and agentic [304] AI systems.		
	Ethical & Safe Use	Trust & Safety	TO.1, ST.6, SI.5
	Design AI systems that promote healthy human-AI interactions and protect a user's mental health [305, 306, 307].		
	Be proactive in designing mechanisms to combat issues of dependence [308], overreliance [309, 310], and emotional attachment [311].		
	Design human-centered interventions targeting the prompting process to promote more responsible and ethical use [312].		

Map

The Map stage captures the activities that provide visibility into the organization and the interconnected web of its external relationships, organizational responsibilities, assets, and data. Maintaining this situational awareness is critical for nearly every other aspect of this guide. Without visibility, it becomes exceedingly difficult—if not impossible—to ensure the proper management, security, and compliance of the organization's operations.

	Map	Owner	Guidance
Organization	**Stakeholders**	Strategy	ST.1, PP.1, IR.1
	Conduct a stakeholder analysis to identify and map the organization's key stakeholders [313, 314].		
	Identify stakeholders directly related to or impacted by the adoption of AI technology [315].		
	Group and prioritize stakeholders using a power-interest matrix, salience model, or similar technique [316, 317]		
	Solicit input from stakeholders throughout the AI adoption process (see Engage Stakeholders).		
Operations	**Supply Chain**	Supply Chain	IT.2, SC.2, RR.3, IV.2
	Conduct a supply chain mapping to identify the external dependencies of the organization [318, 319, 320].		
	Identify potential threats and risks to the AI supply chain [321], in particular those related to data used by AI systems [322].		
	Use the supply chain map to inform the development of a supply chain risk management plan [323].		
Workforce	**Ownership**	Leadership, Operations, Workforce	IT.3, IV.5, RC.1, SL.2, WF.2, IS.9, PP.3, PP.8, PI.1, MO.8, IR.1, IR.2
	Define who owns the AI system, components, model, and associated data within the organization [324].		
	Strategically position AI leaders within the organization's reporting structure and ensure they have the needed authority [325].		
	Map AI-related responsibilities [326, 327] to specific functional units, teams, and individuals (use the mapping in this report as a guide).		
	Delegate responsibilities and authority to individual teams [328] while also providing formal structures to support them [329].		
	Record ownership of organizational assets and related responsibilities in the inventory (see Map Assets).		

	Map	Owner	Guidance
Workforce	**Approval**	Leadership, Operations	IT.6, MG.4, DD.7, DD.8
	Establish a documented approval process and define where in the life cycle approvals are required (see Plan Life Cycle).		
	Map and record the approval workflow (i.e., who is approving, when, in what order) at each stage and for each system component [330].		
Infrastructure	**Assets**	Infrastructure	IV.1, IV.2
	Create a centralized inventory for tracking IT assets. Include the system, components, model, and data in the inventory [331].		
	Use established templates [332] or, preferably, a more advanced asset management systems to track AI-related assets [333].		
	Include relevant metadata about the asset, including provenance (see Map Provenance) and ownership (see Map Ownership).		
	Data Flows	Infrastructure	IV.3, TR.4, SM.7, PP.8
	Conduct a data flow mapping to identify how data moves through the organization, how it is processed, and where it is stored [334, 335].		
	Create a visual representation of the organization's data flows using a data flow diagram [336].		
	Use the data flow mapping to assess privacy risks (see Assess Data Protection) and ensure privacy compliance (see Assess Legality).		
Data	**Classification**	Privacy, Data	IS.1, IV.4
	Systematically classify assets [337] (see Define Classification Scheme) and determine their value in terms of criticality [338].		
	Consider methods to automate the assessment and classification of organizational assets [339].		
	Use the classification and valuation to define the safety, security, and privacy controls required to protect the asset [340].		
	Provenance	Data	FS.5, TR.1, TR.4, IS.7, PI.10
	Trace licenses, creators, uses, and sources to generate standardized documentation regarding dataset lineage [341].		
	Use cryptographic tools such as digital signatures and watermarking to map content and data provenance [342].		
	Replace static reports with interactive provenance visualizations that trace the full AI data life cycle [343].		
Model	**Traceability**	Innovation, Product	TR.2, TR.3, TR.4, TR.6, TR.7
	Embed watermarks using semantic-preserving code transformations so that AI-generated outputs remain traceable [344].		
	Map model traceability using standardized metadata, secure registries, and immutable logs [345].		
	Implement mechanisms to track and record the activity of systems acting autonomously, such as AI agents [346].		

	Map	Owner	Guidance
System	**Components**	Product	IV.1, SC.2, ES.1, TO.6
	Use a bill of materials (BOM) to document system components, including hardware [347], software [348, 349], and AI BOMs [350].		
	Document the AI system and its components, training, evaluation, and intended use through model [351] or use case [352] cards.		
	Consider including information in the system documentation related to the OECD AI classification framework [147].		

Assess

The Assess stage marks the end of the planning phase, before the organization moves on to implementation. This stage involves assessing the plans for the AI system and the organization's capability to implement them safely. It is critical that the organization's leadership review the various assessments below holistically and determine whether it should move forward in adopting the AI system or whether further iteration in the planning phase is required.

	Assess	Owner	Guidance
Organization	**Points of Failure**	Leadership	IM.1
	Identify existing structures and practices that might serve as roadblocks for AI adoption initiatives [353].		
	Understand the technical, and more often social or organizational, reasons why many AI projects fail [354, 355].		
	Incorporate lessons learned from previous AI projects and failures, both internal and external (see Improve) [356].		
	Threats	Cybersecurity	MO.2, DD.2
	Maintain awareness of the cybersecurity threat landscape [357, 358]. Share threat information (see Monitor Information Sharing).		
	Conduct cybersecurity threat assessments to identify potential threat actors [359, 360], leveraging cyber threat intelligence [361].		
	Conduct threat modeling to identify potential vulnerabilities and risks [362, 363, 364, 365], including those specific to AI systems [366].		
	Account for the range of AI-specific attacks, including those related to privacy [367] as well as AI-based attacks [368].		
Operations	**Risk**	Risk	RM.4, RM.5, RM.6, SC.3, IM.7, PM.3, IS.9
	Take a systematic approach to identifying risks [369], including risks [370] and harms [371] specifically related to AI systems [372].		
	Account for AI misuse risks, including chemical, biological [373], radiological and nuclear (CBRN) [374] and cyber [375] risks.		
	Conduct risk assessments [376] using established tools and techniques [377]. Assess the risks stemming from AI adoption [378, 379].		
	Use qualitative and quantitative measures of risk where appropriate [380, 381], accounting for uncertainty [382].		
	Establish a process to prioritize risk [383]. Consider using risk matrices and other prioritization tools to assist in this process [384, 385].		

	Assess	Owner	Guidance
Operations	**Legality**	Legal, Compliance	TO.5
	Assess whether the AI system conforms to applicable laws and regulation (see Define Context), including the EU AI Act [386].		
	Consider the potential liability of the organization for AI-related harms and the applicability of U.S. tort law [387].		
	Evaluate how privacy regulation, such as the General Data Protection Regulation, may impact data use by the AI system [388].		
	Consider applicable antidiscrimination laws [389] and potential civil rights impacts related to the AI adoption and use [390].		
	Ethics	Responsibility, Compliance	TO.5, IM.5, IM.6
	Conduct regular human rights impacts and due diligence assessments of the organization's practices [391].		
	Assess the alignment of the AI system with the organization's mission and ethical principles [253].		
	Implement a multidisciplinary evaluation framework to measure social impacts alongside technical performance [392].		
	Assess ethics through audits that invite third-party evaluation and take into account value sensitive design approaches [393].		
	Supply Chain	Supply Chain	SC.7
	Understand the threats and risks to the organization's supply chain [394] based on the supply chain map (see Map Supply Chain).		
	Conduct a supply chain risk assessment to evaluate risks and dependencies [395, 396], particularly for upstream AI providers.		
	Consider the risks of using open source models, software, or libraries in system development and its future maintenance [397].		
Workforce	**Capability**	Workforce	WF.1, ST.5, DD.1
	Develop or use a maturity grid/model to assess organizational capabilities and how they change over time [398].		
	Assess cybersecurity [399, 400] and AI [401] competencies within the organization's workforce.		
	Determine whether the organization's existing workforce is ready for and can adequately support AI adoption [402].		
Infrastructure	**Platforms**	Infrastructure	DD.3
	Assess the computing needs for AI workloads, which often require specialized hardware for intensive computation [403, 404].		
	Evaluate AI infrastructure options [405], including on-premises [406], cloud [407], and hybrid [408] deployments.		
	Assess the risks [409] and potential security threats [410] of related to the use of cloud computing platforms.		
	Compare the range of AI platform providers and tools if planning on using cloud services or AI-as-a-Service (AIaaS) [59].		

	Assess	Owner	Guidance
Data	**Data Protection**	Cybersecurity, Privacy, Data	PI.7, MS.9
	Use the data mapping (see Map Data Flows) to conduct a data protection [411] and privacy impact assessment [412, 413, 414, 415].		
	Understand the range of privacy harms [416] and the privacy risks that can occur across the AI life cycle [417].		
	Assess the technical and procedural protections available for securing data used by AI systems [418].		
Model	**Technology Readiness**	Innovation	DD.1, IT.2, TO.5
	Assess the maturity of the given AI technology using technology readiness levels [419, 420] applied to AI systems [421].		
	Understand the factors influencing successful AI adoption within organizations [422, 423].		
	Evaluate the maturity of the organization (see Self-Assess) and conduct an AI readiness assessment [424].		
	Trade-Offs	Innovation	IM.7
	Assess the design trade-offs between AI model performance and explainability [425, 426], robustness, and fairness [427].		
	Weigh the benefits of data minimization of sensitive attributes and the limitations when conducting bias assessments [428, 429].		
	Assess the benefits and risks of developing or deploying AI models with dual-use capabilities [430].		
System	**Impact**	Trust & Safety	IM.1, RC.2, SI.3, IT.10, IM.2, IM.5, IM.7, PM.3
	Conduct AI [431, 432, 433, 434] and algorithmic [435] impact assessments, accounting for the system's intended deployment.		
	Assess the impact that the intended AI use case may have on human rights and democratic processes [436, 437, 438].		
	Use established templates for conducting and documenting impact assessments [439, 440, 441].		
	Employ sets of related metrics rather than a single indicator to ensure that assessments capture the system's effects and dynamics [442].		
	Encourage multi-stakeholder participation in assessments across government, academia, civil society, and the private sector [443].		
	Attack Surface	Cybersecurity	DD.2, MO.2, MO.3
	Understand the attack surface [444] of the system, identifying common points of entry or vulnerability in IT [445] and AI [446] systems.		
	Conduct an attack surface analysis to identify high-risk points that will need to be tested and defended [447, 448].		
	Account for vulnerabilities specific to ML [449], generative AI [450], and agentic [451] systems.		
	Take steps to reduce the attack surface by removing unnecessary functionality, code, access points, and system calls [452, 453].		

Acquire

The Acquire stage represents the start of the implementation phase. The degree to which the recommendations in this section apply largely depends on what aspects of the AI system are being developed internally versus acquired from external vendors. The extent of the acquisition can range from the entire system, such as in AIaaS models, down to the data used for model training or open-source libraries used in software development. This section provides guidance that is broadly applicable to all forms of acquisition as well as tailored recommendations for acquiring third-party datasets, models, or AI systems.

	Acquire	Owner	Guidance
Organization	**Relationships**	Leadership, Supply Chain	SC.1, RC.3, AU.6, SI.1, IR.6
	Build relationships with suppliers that enable transparent communication [454] and effective risk management [455].		
	Create a strategy for stakeholder engagement [456] that promotes meaningful participation from AI stakeholders [457, 458]		
	Establish relationships with individuals and organizations in academia [459, 460], civil society [461], and the government [462].		
Operations	**Due Diligence**	Supply Chain, Legal	RC.2, SC.5, TE.4
	Vet potential suppliers using a comprehensive due diligence assessment [463, 464].		
	Quantify the costs, benefits, and risks of doing business with a given supplier as a part of the due diligence process [465].		
	Ensure that suppliers' labor practices are legal and ethical [466], especially for AI-related data annotation and content moderation [467].		
	Select suppliers using a structured evaluation approach, such as multi-criteria decision-making [468].		
	Contracts	Legal, Supply Chain	RC.3, SC.6, VN.2, MS.5
	Closely evaluate the contracts and terms of use involved in the procurement and use of third-party AI models and tools [469, 470].		
	Build key provisions into supplier contracts that address privacy, security, intellectual property, and liability, among others [471, 472].		
Workforce	**External Expertise**	Workforce	ST.4, IM.3, TE.7
	Consider bringing in external AI expertise to augment talent gaps within the organization (see Define Needs).		
	Engage external experts in the AI TEVV and red-teaming processes, particularly when testing for domain-specific capabilities [473].		
	Consider the role that the broader public can play in testing and AI red-teaming activities [474].		

	Acquire	Owner	Guidance
Infrastructure	**Verification**	Supply Chain, Cybersecurity	NS.4, IS.3, ES.4
	Ensure products and services are acquired from trusted and vetted sources only (see Assess Supply Chain, Acquire Due Diligence) [475].		
	Verify the source and integrity of acquired hardware, software, models, data, and other components [476].		
	Take steps to secure the organization's AI supply chain, accounting for the differences between AI and traditional software [477].		
Data	**Consent**	Data, Privacy	FS.5, PP.5, PI.3, PI.4, IS.8
	Obtain informed consent from individuals prior to the collection of personal information and confirm that consent over time [478, 479].		
	Disclose the intended use of collected data, particularly for AI uses [480], and obtain a data processing agreement [481].		
	Provide notice to individuals when their data is being collected [482] and when the policies or use of that data changes [483].		
	Obtain consent from and provide compensation to creative workers whose data is used to train AI models [484, 485].		
	Data Collection	Data	PP.6, PI.2, SG.2
	Determine the appropriate tools and techniques to use for data acquisition, augmentation, and labeling [486].		
	Adhere to the principle of data minimization when collecting and processing personal data (see Promote Data Minimization).		
	Assess ethical, legal, privacy, and data quality issues related to large-scale web scraping used for AI training [487, 488, 489, 490, 491].		
	Scrutinize collected data for potential data poisoning attacks (see Protect Integrity) [492].		
	Third-Party Datasets	Data, Supply Chain	SG.5, SG.2, FS.1
	Assess the various sources of data and the trade-offs in using them for training AI models [493].		
	Follow data supply chain security best practices when acquiring and using third-party data for AI training [494].		
	Weigh the advantages and disadvantages of using synthetic data to train models [495] and the risk of model collapse [496, 497, 498].		

	Acquire	Owner	Guidance
Model	**Third-Party Models**	Innovation, Supply Chain	SG.5, TE.3, TE.4
	Follow relevant public- [499] or private- [500] sector guidance on the strategy and steps to be taken during AI procurement.		
	Scrutinize third-party models for security risks and vulnerabilities that they might introduce (see Map Supply Chain) [501, 502].		
	Assess the challenges in using pretrained models for downstream development [503] and potential dependence on upstream providers.		
	Weigh performance [504], cost [505], documentation [506], transparency [507], and security [508] trade-offs for open models.		
	Determine whether and how the organization will modify the pre-trained model [509]. Assess trade-offs [510] and risks [511].		
System	**Third-Party Systems**	Product, Supply Chain	SG.5, TE.3, TE.4
	Apply a structured approach to procurement and acquisition that includes evaluation and acceptance testing [512].		
	Explore the wide range of AI tools, services, and platforms that are available to identify those that meet the organization's needs [513].		
	Compare different AI offerings [514] and evaluate provider policies for data collection and monitoring. Test models side by side [515].		
	Evaluate potential AI systems against prespecified requirements (see Define System Requirements) in a systemized way [516].		

Develop

The Develop stage complements the previous Acquire stage, covering the development of AI capabilities in-house as opposed to those being acquired from outside sources. While only a handful of companies are likely to be developing their own frontier models from scratch, many are likely to develop their own more specialized models or build on existing, pre-trained models to meet the organization's specific needs. This stage covers the practices and techniques needed for responsible AI development and innovation.

	Develop	Owner	Guidance
Organization	**Responsible Innovation**	Innovation	SI.1, SI.2, SI.3, SI.5, SI.6, IM.3
	Ensure that the organization adheres to responsible innovation practices in its development and adoption of new technology [517].		
	Apply responsible innovation practices to the research and development of AI systems [518, 519].		
	Develop context-specific tools and methods to prioritize fairness, transparency, and accountability in innovation [520].		
Operations	**Management Systems**	Infrastructure	MG.7, IV.4, DD.9, DD.10
	Implement software project management [521] and AI project management [522] best practices.		
	Create a management program to oversee MLOps [523].		
	Establish an asset management system [524, 525] that includes AI- and ML-based assets [333] (see Map Assets and Manage Assets).		
	Develop a dependencies management [526] (see Map Supply Chain) and assurance [527] (see Acquire Due Diligence) program.		
	Put in place common configuration [528] and control [529] management systems (see Promote Common Architecture).		
	Create a risk management function [530] (see Assess Risk and Control Risk), that includes supply chain risk management [531].		
	Develop a vulnerability management program (see Manage Vulnerabilities) [532].		
	Implement change [533] and patch [534] management systems (see Develop Version Control and Develop Patches).		
	Establish an incident [535] and continuity [536] management system (see Respond and Plan Response).		
	Develop a training and awareness program for personnel (see Deploy Training & Drills and Manage Personnel Awareness) [537].		
	Create an overarching situational awareness program within the organization (see Manage Situational Awareness) [538].		

	Develop	Owner	Guidance
Workforce	**Diverse Teams**	Workforce	ST.4
	Build teams that reflect a diversity of backgrounds, knowledge, expertise, and experience [539, 540].		
	Understand the importance of diverse teams for AI development, testing, and operation [541].		
	Develop a recruitment plan to address the scarcity of AI skills, equity gaps in AI talent [542], and lack of diversity on AI teams [543].		
	Facilitate an environment that promotes communication, trust, and psychological safety among individuals on diverse teams [544, 545].		
Infrastructure	**Logging**	Infrastructure	LG.1, LG.2, LG.6, NS.7
	Establish practices to collect, store in a centralized location (see Monitor Centralized Analysis), manage, and protect event logs [546].		
	Prioritize logging within the organization based on the criticality of systems and data sources [547].		
	Determine what events to log, what metadata to capture, and how to store them using standardized log formats [548].		
	Ensure that logs capture the right information to meet transparency [549], auditing, and issue investigation needs [550, 551].		
	Log information and events from AI systems that capture a range of responsible AI metrics, not just model performance [552, 553].		
	Version Control	Infrastructure	MG.7, TR.1, TR.2
	Use a version control system to manage system artifacts, coordinate development, and control changes [554].		
	Apply version control principles to AI and ML workflows to ensure reproducibility in experiments, development, and testing [555].		
Data	**Data Quality**	Data	TR.5, SG.2, IS.4
	Identify the data quality requirements that are important across different stages of AI development [556].		
	Adopt an ML-focused, but model-agnostic, approach to data quality management [557].		
	Employ techniques to assess and improve the quality of data within the organization [558].		
	Treat data as a governance mechanism, managing and improving data across the supply chain to improve AI safety and security [559].		
Model	**Alignment**	Innovation	IT.7, TO.5
	Work with internal and external stakeholders to identify the values, goals, and behaviors with which to align the AI system [560].		
	Align the behavior of the model using forward alignment techniques during training and backward alignment post-training [561].		
	Evaluate the model for deceptive, scheming, or other misaligned behaviors [562].		

	Develop	Owner	Guidance
Model	**Robustness**	Innovation	IM.4, SG.1, SG.2
	Apply robustness techniques to training data, architecture design, model training, and post-processing [563].		
	Include adversarial examples in model training to help make models more robust to adversarial attacks [564].		
	Consider including a consistency alignment training step to improve the robustness of model responses [565].		
	Use robust training frameworks to combat data heterogeneity and Byzantine attacks if using federated learning [566].		
	Employ techniques such as decision-boundary based federated adversarial training for robust federated learning [567].		
	Transparency	Innovation	TO.1, TO.2, TO.6
	Implement mechanisms and employ tools to help provide interpretations of what led models to produce their outputs [568].		
	Incorporate explainability features into the system that support users' understanding the AI model and its output [569].		
	Develop a use-case-centric explainability framework that tailors model response explanations to users and decision contexts [570].		
	Combine model output with symbolic reasoning to improve the transparency of automated decision-making processes [571].		
	Federated Learning	Innovation	PP.7
	Consider federated learning for training ML systems, especially in contexts where data is heterogeneous and unsafe to centralize [572].		
	Use federated learning to share or migrate models securely in cases where local training data is insufficient [573].		
	Acknowledge that federated learning is not a one-step privacy solution and that it may still present security risks [574].		
	Understand the additional risks of and potential attacks on a federated learning system [575].		
	Privacy Preservation	Innovation	PP.7
	Weigh the privacy guarantees provided by differential privacy against utility and performance trade-offs [576].		
	Consider using federated learning (see Develop Federated Learning) to help preserve the privacy of personal data used in training [577].		
	Assess how to minimize the use of sensitive data and features at inference time, while retaining performance [578].		
System	**Secure Software**	Innovation, Product	DD.1, DD.4, VN.2
	Ensure that the organization implements and follows secure software development practices [579].		
	Take steps to mitigate risks in software development, the supply of third-party components, and the build and delivery of software [512]		
	Apply security best practices to the development of AI models and the related code that make up the AI system [580, 581].		

	Develop	Owner	Guidance
System	**Patches**	Innovation, Product	VN.5, ES.8
	Understand the security patch landscape [582] and the challenges in applying patches within an organization [583, 584].		
	Implement a proactive approach to patch management that includes proactive testing, prioritization, and monitoring [585].		
	Consider incorporating automation at various stages in the patch management process [586].		

Control

The Control stage captures the practices and technical solutions to mitigate the risks that may result from the organization's practices, the activity of its personnel, and the operation of its systems. In general, this stage addresses the potential harm stemming from the organization, while the following Protect stage addresses the potential threats to the organization. Not all of these harms necessarily originate from within the organization or result from unintended action; however, many of these risks do.

	Control	Owner	Guidance
Organization	**Physical Environment**	Physical Security	SM.5, SM.6, PS.1, PS.2, PS.3, IA.4, PS.5
	Allocate personnel and security equipment to maintain physical control over the organization's facilities [587].		
	Control physical access to facilities and restricted areas, including the use of photographic, video, and radio-frequency devices [588].		
	Employ controls that help prevent environmental hazards such as fire, water, electrical, and natural disasters [589].		
Operations	**Risk**	Risk	RM.7, RM.8, IM.4, SC.3
	Determine whether the organization will reduce, transfer, avoid, or accept identified risks (see Assess Risk).		
	Implement mitigation strategies to reduce the organization's IT [590] and AI [591, 592] risks.		
	Consider the use of insurance to transfer some of the risk related to AI from the organization to providers [593].		
	Identify opportunities for risk avoidance when adopting new technology [594]. Do not pursue AI if the risks are too great [595, 596].		
	Determine whether to accept risk if it meets defined acceptance criteria [597] and risk tolerance (see Define Requirements).		
Workforce	**Activity**	Workforce	MS.4, MS.5, MS.6, MS.10, NS.2, PS.2
	Screen potential employees and conduct background checks during hiring [598, 599] to reduce the potential of insider threats [600].		
	Establish an internal control system to guide personnel behavior and enforce compliance policies, practices, and laws [601, 602].		
	Employ a separation or segregation of duties policy that prevents full control of a process or asset by any one individual [603].		
	Revoke access, ensure the return of organizational assets, and otherwise mitigate risk during employee separation [604].		

	Control	Owner	Guidance
Infrastructure	**Identities**	Infrastructure	IA.1, IA.3, IA.5, IA.9
	Establish centralized management of identities, credentials, and identity proofing within the organization [605, 606, 607].		
	Enforce a secure password policy and promote the use of password managers [608, 609].		
	Use stronger authentication mechanisms such as multifactor authentication [610] and single sign-on where possible [611].		
	Extend identity management to agentic AI, accounting for the scale, complexity, and dynamic and ephemeral nature of agents [612].		
	Access	Infrastructure, Cybersecurity	DD.5, IA.4, IA.6, IA.7, AC.3, AC.6, AC.7, AC.8, NS.5, MS.6, PP.9, LG.5
	Restrict access of personnel and systems to only what is strictly necessary (see Promote Least Privilege & Functionality) [613].		
	Select and implement an access control model that is appropriate for the organization's maturity, size, and requirements [614, 615].		
	Control what assets personnel have access to, particularly the ability to modify components of the AI system [616].		
	Control what assets the AI system has access to, including what information it can ingest and what actions it can take [617].		
	Data Flows	Infrastructure, Data	SM.7, IA.5, NS.2, NS.3, NS.5, IS.4, MS.2, MS.7, DD.6
	Implement network segmentation (see Protect Network) and enforce access and data flow restrictions across segments [618].		
	Employ controls, such as firewalls [619] and cross-domain solutions [620], at domain boundaries to control the flow of information.		
	Restrict the flow of data across security domains based on its confidentiality classification (see Define Classification Scheme).		
Data	**Data Transfers**	Data	PI.6, PI.9, SM.7, MS.7
	Limit the transfer of personal data outside of the organization to what is strictly necessary (see Promote Data Minimization).		
	Ensure international data transfers comply with relevant privacy legislation [621, 622, 623, 624, 625, 626].		
	Implement controls to prevent the transfer of data via insecure or inappropriate channels that may lead to data leakage [627].		
	Extend these controls to data sent by personnel to commercial AI products via web interface, file upload, and other protocols [628, 629].		

	Control	Owner	Guidance
	Classified & Proprietary Data	Data, Privacy	SM.7, DD.5
	Evaluate the range of proprietary, sensitive, or classified information that employees might send to external AI services [630, 631].		
	Consider that these risks will exist whether the organization provides AI tools or employees use their own personal services [632].		
	Establish policies for acceptable AI use (see Plan Policies) and train employees on responsible use (see Deploy Training & Drills).		
Model	**Bias**	Innovation, Product	FS.1, FS.3, FS.4
	Adopt practices to assess and mitigate issues of equity, diversity, and inclusion related to algorithms and AI use [633, 634].		
	Implement a bias mitigation system that enables dynamic adjustments to model outputs that mitigate bias without retraining [635].		
	Use causal models to detect and correct poorly weighted relationships between sensitive data attributes that lead to bias [636].		
	Implement guardrails to steer model behavior toward unbiased output [637].		
	Inputs	Innovation, Product	IS.4, SG.3, SG.6
	Validate and sanitize inputs to the system [638]. Extend these methods to AI models to handle harmful or malicious input [639].		
	Remove or mask privacy-sensitive information that users provide to AI models [640].		
	Employ controls to prevent injection-based attacks [641], including injection attacks on generative AI models [642, 643, 644].		
	Outputs	Innovation, Product	IS.4
	Selectively prune training data, fine-tune the model, or develop output controls to reduce the likelihood of harmful outputs [645].		
	Build guardrails into the AI system to flag and handle model output that may be inaccurate [646] or harmful [647, 648, 649].		
	Consider that harm may only occur in aggregate or across multiple outputs of the model [650].		
	Synthetic Media	Trust & Safety	FS.5
	Follow best practices in the development, creation, and distribution of AI-generated synthetic media [651].		
	Deploy tools and processes within the organization to identify and protect against the use of deepfakes [652].		
	Apply watermarking techniques to AI-generated media to help identify the content as synthetic [653].		
	Employ content provenance solutions to strengthen the verifiability of non-synthetic media [654].		
	Identify opportunities for using synthetic data in the AI development pipeline [655] and assess the related risks [656].		

	Control	Owner	Guidance
System	**Action Space**	Product, Infrastructure	AC.2, IS.8, ES.3
	Constrain the set of actions that agentic systems can take and require human approval before more high-cost actions are taken [657].		
	Ensure safe exploration of the action space for AI use cases that employ reinforcement learning techniques [658].		
	Agency	Product	TO.3, SG.1
	Consider the potential harms [659, 660] and greater security risks [661, 662] of creating or deploying systems with greater agency.		
	Deploy agentic systems in narrower environments that are more conducive to their limitations [663] and reliability problems [664].		
	Implement technical and nontechnical guardrails to control and monitor the agency with which the system operates [665].		
	Employ human oversight of systems that operate with any degree of autonomy and protect those channels from compromise [666].		
	Extend access, authorization, and delegation controls (see Control Access) to agentic systems [667].		

Protect

The Protect stage captures the practices and technical solutions to mitigate the range of risks to the organization, its personnel, and its systems. In general, this stage addresses the potential threats to the organization, while the preceding Control stage addresses the potential harms stemming from the organization. Not all of these threats necessarily originate from outside the organization or result from purposeful malicious action; however, many of these risks do.

	Protect	Owner	Guidance
Organization	**Physical Premises**	Physical Security	SM.5, PS.4, PS.5
	Implement physical perimeters, external and internal, that are equipped with sensors, alarms, and entry control [668].		
	Consider additional security protections if the organization houses its own computing resources or data centers [669].		
	Employ automated protection mechanisms in electronic systems to ensure physical safety [670].		
Operations	**Intellectual Property**	Legal	MS.8
	Protect intellectual property (IP) when training AI [671, 672]. Ambiguity exists [673], but a recent case indicates limits to fair use [674].		
	Use digital replicas—that is, AI-generated likenesses of people—only with permission [675] and in a legal manner [676, 677, 678].		
	Protect and respect the IP of AI-generated creations [679, 680], particularly where human contributions constitute authorship [681].		
	Protect the organization's AI-based, and AI-assisted [682], IP through legal [683] and technical means (see Protect Weights).		
Workforce	**Personnel**	Workforce, Compliance	RC.2, RC.4, PS.5, MO.8
	Protect the rights and well-being of employees [684], particularly as the organization operates in an increasingly AI-driven world [685].		
	Ensure the organization, and its suppliers, responsibly employ data workers [686, 687] and other contractors or gig workers [688, 689].		
	Help protect the mental health of employees exposed to harmful content, such as AI testers [690] and content moderators [691].		
	Protect employees' privacy and well-being if deploying AI tools internally [692], particularly for surveillance purposes [693, 694, 695].		
	Protect employee data, ensure nondiscrimination, and provide transparency if using AI in hiring or promotion processes [696, 697].		

	Protect	Owner	Guidance
Infrastructure	**Network**	Cybersecurity	NS.1, NS.5, NS.6, IS.2, ES.6, MO.6
	Implement network security best practices [698] and account for shifts in the enterprise IT landscape (cloud, microservices, etc.) [699].		
	Separate the network into segments with different access and security, providing defense-in-depth against intruders [700].		
	Employ intrusion detection and prevention systems at external and internal segment boundaries [701, 702].		
	Deploy honeypot and other network deception and obfuscation techniques to divert attackers and gain insights into their behavior [703].		
	Ensure that edge devices are secured and configured correctly [704].		
	Assets	Cybersecurity	IV.4, SM.2, SM.3, IS.2, ES.5
	Deploy and manage assets with secure configurations and up-to-date software [705], including mobile devices [706] and servers [707].		
	Employ antimalware protections [708] and other preventive security tools and services [709, 710].		
	Implement human-centric protections in addition to technical ones to help address the challenge of human factors in cybersecurity [711].		
	Ensure that personnel understand and follow basic cybersecurity practices to protect their devices [712].		
	Logs	Infrastructure	LG.5, LG.6, LG.7, MG.6
	Protect logging and audit mechanisms from tampering and accidental failures [713].		
	Restrict access to, encrypt, and create redundant backups of log data. Store log data separately from operational systems [714].		
	Backups	Infrastructure	RR.5, RR.7
	Create consistent backups of critical data and systems for restoration in case of a failure of compromise [715, 716].		
	Evaluate on-site, off-site, and cloud storage options for maintaining backups. Retain multiple copies in different locations [717].		
	Maintain and test backups to help protect against ransomware and other data loss events [718].		

	Protect	Owner	Guidance
Data	**Confidentiality**	Privacy, Cybersecurity	IS.2, IS.5, PI.7, PP.7, PP.8
	Defend against common data breach vectors to protect the confidentiality of data [719].		
	Establish safeguards to protect PII and minimize exposure in the case of a data breach [720].		
	Use established post-quantum encryption [721] and key-exchange standards [722] to encrypt data.		
	Remove classified, sensitive, and proprietary information from training datasets and monitor model output for data leakage [723].		
	Protect models against membership inference attacks that can leak information about the underlying training dataset [724, 725].		
	Defend against model inversion attacks that can be used to extract information about the model and its training data [726, 727].		
	Integrity	Cybersecurity	IS.3, ES.3
	Use post-quantum digital signature standards [728, 729] to verify the integrity of data and the identities of senders and receivers [730].		
	Employ integrity-checking mechanisms to validate data, ensure consistency, and detect anomalies [731, 732].		
	Protect the integrity of data used to train and operate AI systems from sourcing to storage to decommissioning [733].		
	Defend against a range of data poisoning attacks that can be used to manipulate AI model behavior [734, 735, 736, 737].		
Model	**Weights**	Cybersecurity	IS.2, IS.5, SM.4
	Conduct a threat assessment and implement a security plan to prevent the theft of model weights [738, 739, 740].		
	Implement protections against unauthorized knowledge distillation and model extraction [741, 742].		
	Consider the risks of model weights leaking through side-channel attacks when deployed on untrusted hardware [743, 744].		
	Consider using trusted execution environments (TEEs) to protect model weights running on untrusted hardware (see Protect Execution).		
	Protect against model poisoning attacks, whether through data poisoning (see Protect Integrity) or direct manipulation [745].		
	Inputs & Outputs	Cybersecurity	SG.3, IS.2, SM.4, MS.3
	Use encryption and secure communication channels to ensure the integrity of model input and output [746].		
	Defend against a wide range of adversarial attacks [747, 748, 749], including model evasion [750, 751] and manipulation [752].		
	Account for multi-turn attacks that can occur across multiple sets of inputs or interactions [753].		
	Consider the risk of exposing model vectors and embeddings [754], as this can leak information about the model and data [755, 756].		
	Defend against side-channel attacks that can be used to reconstruct model input and output [757, 758, 759].		

	Protect	Owner	Guidance
Model	**Execution**	Infrastructure, Cybersecurity	IS.3, ES.4
	Assess security and performance trade-offs [760] of using TEEs to protect model execution on untrusted hardware [761, 762].		
	Stay informed of advances in and limitations of secure computing for GPU [763, 764], NPU [765], TPU [766], and DPU [767] hardware.		
	Protect the model from revealing, or being able to reveal [768], the system prompts guiding the model's behavior [769, 770].		
	Resource Utilization	Infrastructure	NS.6, NS.8, RR.4
	Limit or throttle the volume of requests that any given user can make to the model to prevent resource exhaustion [771].		
	Protect against economic denial of sustainability (EDoS) attacks on cloud computing resources used by the organization [772].		
	Implement protections to identify and mitigate similar resource consumption attacks on AI models [773, 774, 775].		
System	**Availability**	Cybersecurity, Infrastructure	NS.6, RR.4
	Conduct capacity planning [776]. Forecast and allocate computing resources [777, 778]. Dynamically adjust to meet demand [779].		
	Allocate adequate resources to maintain availability and protect against resource-based attacks (see Protect Resource Utilization).		
	Implement defenses against denial of service (DoS) attacks on the organization's networks, services, and public-facing APIs [780].		
	Source Code	Cybersecurity	IS.2, ES.2, ES.3
	Use trusted code repositories to maintain code and implement secure repository practices [781].		
	Employ a version control system (see Develop Version Control) to manage changes and prevent unauthorized modification.		
	Separate secrets (e.g., credentials, keys) from source code, block their inclusion in repositories, and regularly scan code for them [782].		
	Protect access to online repositories from adversaries seeking to steal, poison, or extract credentials from source code [783].		
	Users	Trust & Safety	TO.7, IS.4, FS.4, IM.4
	Design AI as a supplement to—not substitute for—human intelligence [784, 785], interaction [786], learning [787], and work [788, 789].		
	Protect users from developing dependence [790] or overreliance [791] on the organization's AI. Avoid addictive design patterns [792].		
	Be hesitant to develop AI systems with anthropomorphic qualities [793]. If used, address the risk posed to users [794, 795, 796].		
	Acknowledge the potential for severe harm if deploying AI technology in a role that serves as—or could be used as—a replacement for important human-to-human interaction, such as in AI assistants [797], therapists [798, 799], or personal companions [305].		
	Account for the use of the organization's AI tools or services by at-risk populations [800, 801], particularly children [802, 803, 804].		

Test

The Test stage, unsurprisingly, comprises the range of testing procedures that should be carried out to evaluate the organization's practices, competencies, infrastructure, and systems. The results of these tests will help inform the decisions made in the Approve stage. While testing at this point in the adoption life cycle is critical, ongoing testing should occur at regular intervals even once the practices or systems are deployed.

	Test	Owner	Guidance
Organization	**Business Continuity**	Operations	IR.1, RR.5, RR.8
	Use tabletop exercises, or other discussion- or practice-based exercises, to evaluate business continuity plans [805].		
	Use resiliency scoring metrics for measuring and tracking the effectiveness of continuity plans over time [806].		
	Consider using AI models to assist in generating scenarios that stress test the organization's business continuity plans [807].		
Operations	**Incident Response**	Operations, Cybersecurity	RR.3, RR.5, RR.8
	Conduct tabletop exercises to play out and evaluate the organization's response [808].		
	Use external resources to develop incident response tabletop exercises that address a variety of potential incidents [809].		
	Apply a quality assurance approach to testing the readiness of the organization to respond to incidents [810].		
Workforce	**Competency**	Workforce	ST.5
	Implement regular evaluations of employees' AI skills to help identify talent and training gaps [811].		
	Tailor competency evaluations and programs to the demonstrated needs of personnel and the organization (see Define Needs) [812].		
	Use validated instruments to measure and evaluate employee AI literacy [813].		
Infrastructure	**Controls**	Cybersecurity	IS.6, TE.4, PM.3
	Conduct cybersecurity penetration testing and red teaming to evaluate defensive capabilities and identify weaknesses [814].		
	Test the organization's monitoring, alerting, and intrusion-detection systems [815].		
	Follow best practices for testing web- and user-facing protections and controls [816].		

	Test	Owner	Guidance
Infrastructure	**Updates**	Innovation, Product	VN.7, ES.8
	Thoroughly test all updates in a sandboxed environment or virtual machine before releasing them to production [817].		
	Verify the authenticity of patches, evaluate the effectiveness of the solution, and test for compatibility issues before deploying [818].		
	Evaluate the patch management against best practices to help streamline the patching and update process [585].		
	Restoration	Infrastructure	RR.5
	Test the complete process of restoring systems and data from stored backups at a regular interval [819].		
	Evaluate data backups and system snapshots periodically to ensure that they are free from compromise [820].		
Data	**Bias & Skew**	Data	FS.2, TR.5, TO.4
	Understand the many sources of potential bias in the AI life cycle and develop tests to measure their influence [371].		
	Test the system and the data used to train it for issues of bias and skew [821]. Embed fairness testing into data management [822].		
	Evaluate whether the training data reflects the diversity of the populations where the model will be used [823].		
	Improve data quality (see Develop Data Quality) and employ fairness-aware preprocessing techniques to minimize bias [824].		
Model	**Capabilities**	Innovation	SG.4, IM.4, TR.3
	Evaluate the capabilities of AI models using established benchmarks and conduct red teaming to assess dual-use risks [825].		
	Adopt a structured framework for testing model capabilities to systematically identify model strengths and weaknesses [826].		
	Test the reasoning capabilities of models using a diverse set of benchmarks and compare to human performance baselines [827].		
	Evaluate the capabilities of systems that act with agency or autonomy [828]. Use appropriate benchmarks for agentic capabilities [829].		
	Performance	Innovation	TE.3, TE.4, TE.8, IT.10, IM.4
	Test the generalization and robustness of models to ensure reliable performance after deployment [830].		
	Employ methods to verify the performance and reliability of AI systems in realistic settings [831].		
	Evaluate the model and its performance along various metrics of trustworthiness [832].		
	Assess model performance and reliability on real-world tasks and likely downstream use cases [833].		

	Test	Owner	Guidance
Model	**Fairness**	Trust & Safety	FS.2, SI.1
	Test model fairness using multiple quantitative metrics chosen based on the ethical and social context of the model's application [834].		
	Test the systems involved in automated decision-making for issues related to equity, fairness, and discrimination [835].		
	Evaluate real-world outcomes and trade-offs for affected groups, rather than solely relying on parity-based metrics [823].		
	Use statistical-comparison methods to detect bias between human and AI-generated outputs when using synthetic training data [836].		
	Consider using synthetic data with bias to evaluate the model's performance under a range of fairness conditions [837].		
	Test the model's explainability and interpretability—key characteristics that have implications for bias, fairness, and trust [838].		
System	**Security**	Cybersecurity	DD.6, TE.4
	Conduct information security assessments of the system [839] and robust application security testing [840].		
	Ensure that the developer is able to verify that the system meets minimum software security standards [841].		
	Employ fuzzing methods to identify vulnerabilities and system failures [842]. Use automated methods for root cause investigation [843].		
	Evaluate how the AI system performs against adversarial attacks [844] and other attacks specific to AI/ML systems [8].		
	Conduct red-teaming exercises to assess the security risks of the AI system [845], particularly for agentic AI systems [846].		
	Safety	Trust & Safety	TO.4, TE.8
	Conduct model and contextual safety evaluations to measure potential safety risks [847].		
	Use a life-cycle-wide evaluation framework that evaluates model and system components for safety [848].		
	Measure model safety against established benchmarks and track performance over time [849].		
	Evaluate the safety of the AI system in the context in which it will be deployed and operate, rather than just the model in isolation [850].		
	Consider using automated tools to generate and expand the scope of test cases and test data [851].		
	Quality	Product	IT.8, DD.7
	Identify key assessment metrics for measuring the quality of software and systems [852].		
	Conduct criteria-based assessments to evaluate whether the system meets requirements (see Define System Requirements) [853].		
	Evaluate the system using quality assurance and validation techniques specific to AI systems [854, 855, 856].		

	Test	Owner	Guidance
System	**Acceptance**	Infrastructure, Product	TE.2, TE.7
	Use risk acceptance criteria to structure AI-specific safety cases for acceptance testing [857].		
	Conduct a range of acceptance tests to ensure the system meets safety, security, compliance, and regulatory requirements [858].		
	Explore automated acceptance testing in contexts that do not rely on user input (i.e., not beta testing) [859].		
	Conduct separate user acceptance testing to ensure the system meets end user requirements [860].		
	User Interaction	Product	ST.6
	Employ human-centered evaluations that involve real end users and assess interactions across high- and low-stakes activities [861].		
	Conduct interactive evaluations to capture how AI systems affect users over sustained use [862].		
	Assess whether AI interactions disproportionately affect certain groups of users [863], particularly those that are vulnerable [864].		
	Evaluate the potential influence that the AI system will have on human decision-making and determine if appropriate [865].		
	Evaluate how users can purposefully misuse the AI system and assess the likelihood of that misuse [866].		
	Employ metrics that capture how operators interact with AI systems when deployed in human-machine teaming situations [867].		

Approve

The Approve stage represents the end of the implementation phase. This section covers the decision of whether to proceed with deployment and the inputs that should be assessed in making that decision. Although this stage in the adoption life cycle represents the most important decision-making point, review and approval should also occur at multiple, predefined points throughout this process. This should include, at minimum, an approval step between the planning and implementation phases.

	Approve	Owner	Guidance
Organization	**Go/No-Go Decision**	Leadership	IT.8, RM.3
	Incorporate inputs, such as test results and readiness criteria, from all relevant teams when making high-level approval decisions [868].		
	Prioritize decision criteria (e.g., user acceptance, performance) in alignment with the organization's goals (see Define Objectives) [869].		
	Construct processes that enable decision-makers to change or abandon previously chosen courses of action [870].		
Operations	**Stakeholder Input**	Communication	ST.3, ST.4
	Gather insights from identified stakeholders (see Map Stakeholders) at key stages throughout the adoption life cycle [871].		
	Create organizational incentives to gather stakeholder input and incorporate it into approval decisions [872].		
	Determine the goal of the stakeholder input process and the extent to which stakeholder input will affect approval decisions [873].		
	Risk Threshold	Risk	PM.2, RM.3
	Determine whether the system exceeds the organization's risk tolerance (see Define Requirements) based on testing results.		
	Consider using prespecified "if-then" commitments to govern decisions on AI risk and how systems will be deployed [874].		
Workforce	**Independent Validation**	Workforce	DD.6, TE.7
	Ensure that the system and testing results are reviewed by the independent review committee (see Define Oversight).		
	Engage external and third-party auditors to review AI models and systems that have been developed in-house [875].		
	Seek out external subject matter experts to support internal AI auditing efforts [876].		

	Approve	Owner	Guidance
Infrastructure	**Integration & Acceptance**	Infrastructure	TE.1, DD.10
	Review and approve the results of system integration and user acceptance testing (see Test Acceptance).		
	Address potential adoption hurdles and factors affecting user acceptance of new IT [877] and AI [878, 879] technology.		
	Foster communities of practice around AI use within the organization and create a center of excellence to help obtain user buy-in [880].		
Data	**Privacy Assurance**	Data	PI.1, PI.10
	Ensure the organization has obtained informed consent for any personal data processed by the system (see Acquire Consent).		
	Establish an assurance mechanism to review that data processing by the system aligns with the organization's privacy policy [881].		
Model	**Release Criteria**	Product	IT.8, DD.10
	Determine whether the model meets the requirements specified in the release criteria (See Define Release Criteria).		
	Assess the model against general AI release-readiness checklists [882].		
	Understand that AI evaluations have limits [883, 884] and that unanticipated risks or emergent capabilities can occur [885, 886].		
System	**System Review**	Product	IT.6, DD.7, PM.2, SG.4, TE.5
	Require the developers of the system to demonstrate that the implementation meets requirements (see Define System Requirements).		
	Establish standard peer reviews of the model, software, and system by individuals external to the development team [887].		
	Conduct internal auditing of the AI system throughout the development life cycle to produce an overall audit report [888].		
	Conduct a management review of the system, including a review of the test results [889].		

Communicate

The Communicate stage marks the start of the deployment phase. This stage involves both the documentation and communication of organizational policies, practices, and technical systems. As part of this stage, the organization must determine what information remains for internal use and what information or data is shared with external audiences or published publicly, balancing the need for transparency with the privacy and security risks.

	Communicate	Owner	Guidance
Organization	**Objectives**	Leadership	SM.1, MG.2, AU.7, IM.2
	Engage with personnel to build internal support for the organization's mission and objectives (see Define Objectives) [890].		
	Be transparent internally about how the goals for AI adoption align with the overarching mission of the organization [518].		
	Present the organization's AI strategy and its objective for AI adoption to stakeholders and work to attain buy-in [891].		
	Policies	Leadership, Communication	AU.7, MG.4, MG.5, RC.1, SL.1, WF.4, SM.1, PP.5
	Describe organizational policies clearly in public- or employee-facing documentation and update these documents regularly [892].		
	Use organizational policies, and their communication, as tools to frame values and shape organizational identity [893].		
	Solicit continuous input and feedback (see Engage Stakeholders, Engage Personnel) on policies and where they can be improved [894].		
Operations	**Practices**	Operations	AU.7, MG.4, MG.5, RC.1, RC.5, ST.2, PP.5, PI.3
	Document procedures to standardize practices and facilitate their communication across the organization [895].		
	Communicate to personnel and stakeholders how the organization's policies and practices relate to its AI strategy [896].		
	Decisions	Leadership, Operations	MG.5, IS.7, PI.10
	Document and communicate the reasoning behind the decisions that the organization makes [897].		
	Assess how AI systems may influence organizational decision-making [898] and provide guidance on how to calibrate AI input [899].		
	Develop practices and processes for clearly explaining decisions made with or by AI systems [900].		

	Communicate	Owner	Guidance
Workforce	**Responsibilities**	Workforce, Communication	IV.5, RM.3, WF.4, SM.1, PP.3
	Communicate roles and responsibilities to personnel for the development, management, and oversight of the AI system [901].		
	Establish clear roles and communication channels between business and analytics (or technical) teams in developing new products [902].		
Infrastructure	**Controls**	Cybersecurity	TO.4, DD.10, RM.10
	Use transparency around internal safety and security efforts to build trust with external stakeholders [903].		
	Balance transparency with the potential of information disclosed about controls and safeguards being used by attackers [904].		
	Where applicable, notify users of acceptable use policies and how they will be enforced (e.g., denials, account bans) [183].		
	Changes	Product, Cybersecurity	VN.1, TR.1, TR.2
	Communicate changes to models, as they can be difficult to detect in downstream uses while also having important impacts [905].		
	Provide end users with clear and informative alerts when updates or security patches for the system are available or implemented [906].		
	Disclose AI-related vulnerabilities and risks to downstream parties once identified and when mitigations have been implemented [907].		
Data	**Data Practices**	Data, Privacy	TR.3, PP.5, PI.3, PI.9, PI.10
	Provide clear, easy-to-read documentation of the organization's data and privacy practices in a publicly posted privacy policy [908].		
	Disclose data protection practices to stakeholders through interactive interfaces for increased accessibility and usability [909].		
	Document and communicate data practices for, or involving the use of, AI systems [910].		
	Data Transparency	Data	FS.3, PP.5
	Clearly document the use of upstream datasets in model training where possible [911].		
	Publish a training data declaration to provide transparency around the data used to train or fine-tune the AI model [912, 913].		
	Provide ample documentation—and account for privacy risks—if providing open datasets for public use [914].		
Model	**Risks & Limitations**	Risk, Communication	TO.1, IM.5, FS.3, RC.5, RM.10
	Communicate known risks to stakeholders and interested parties [915]. Consider publishing an organizational risk profile [916].		
	Document the model and its development to provide transparency around its risks, capabilities, and limitations [917, 918].		
	Publish a pre-deployment AI risk disclosure to inform downstream use of the model [919].		
	Maintain and publish up-to-date safety-critical information related to the model [920]. Regularly publish transparency reports [921].		

	Communicate	Owner	Guidance
System	**Test Results**	Innovation, Product, Communication	TE.3, TE.6, TR.3, TR.6, SG.4, RM.10
	Document the process, activities, and results of testing to enable comparison across models and organizations [922, 923].		
	Communicate test results to the development team and ensure that actions are taken to rectify issues identified during testing [924].		
	Publish test results—to the extent that doing so does not compromise the safety or security of the system—publicly to build trust [925].		

Deploy

The Deploy stage consists of the release of the AI model or deployment of the AI system, internally or publicly. This process includes the integration of the system into existing infrastructure, training personnel to support it, and adding system support functions to the organization's day-to-day operations.

	Deploy	Owner	Guidance
Organization	**Review Cycle**	Leadership, Compliance	SL.3, MG.8, AU.3
	Integrate the deployed system and related procedures into the regular management review cycle within the organization [926].		
	Apply a continuous improvement strategy focusing on incremental change and improvement to the system and processes [927, 928].		
	Conduct regular audits of the AI system including related data, controls, risk mitigations, and operating procedures [929, 930, 931].		
Operations	**Enforcement**	Compliance, Workforce	AU.4, SL.1, WF.4, SM.1, DD.10, AC.8, MS.5, MS.7
	Understand the various factors that influence compliance with organizational safety and security policies [932, 933].		
	Evaluate the range of internal compliance strategies and tactics to determine the approach that best suits the organization [934, 935].		
	Assess potential harms to employees before deploying surveillance tools [694, 936], and be transparent about their use [937].		
Workforce	**Training & Drills**	Workforce	SC.4, SI.3, WF.1, IA.8, ES.5, IR.1, IR.2, RR.8, TO.1
	Provide training to aid personnel in improving their AI skills [938] and help meet organizational talent needs (see Define Needs).		
	Consider deploying a rapid occupational training program based on an existing model to develop more advanced AI talent [939].		
	Collaborate with external partners [940] and scale up internal programs [941, 942, 943] to further the availability of AI training.		
	Conduct tabletop exercises [944] and drills [945, 946, 947] to improve personnel's response capabilities and to test response plans.		
Infrastructure	**Architecture**	Infrastructure	IT.4, MO.7, LG.1
	Develop an enterprise integration architecture to ensure the interoperability across the organization's IT infrastructure [948].		
	Conduct early integration planning, compatibility specification, and integration testing to facilitate system deployment [949].		
	Integrate the system [950] and its associated data and events [951] (i.e., for logging and monitoring) into the enterprise architecture.		

	Deploy	Owner	Guidance
Infrastructure	**Inventory & Mappings**	Infrastructure	IV.1, IV.2, IV.3, SM.8
	Add the system and its components to the asset inventory (see Map Assets).		
	Update the data flow mapping to include the new system and how it consumes, produces, and processes data (see Map Data Flows).		
	Employ tools to facilitate the automatic detection of IT assets and the automated updating of inventory and mappings [952, 953].		
Data	**Disclosure**	Privacy, Legal	PI.9, IS.9
	Do not disclose or transfer personal data without consent (see Control Data Transfers).		
	Use differential privacy techniques to provide transparency around datasets and models while managing privacy risks [954, 955].		
	Understand that public datasets containing personal data present privacy risks [956], even if anonymization steps are taken [957].		
	Review the legal and privacy implications of disclosing information about, or the weights of, a model trained on personal data [958, 959].		
Model	**Release**	Product, Infrastructure	IT.9, DD.8, IS.9
	Ensure the model meets the prespecified requirements for release (see Define Release Criteria).		
	Implement the organization's release plan (see Plan Release), making adjustments based on testing results as necessary.		
	Determine the appropriate transparency and due diligence needed based on the type of model and degree of openness [960].		
System	**Deployment**	Product, Infrastructure	IT.9, DD.8, IS.9
	Ensure the system meets the prespecified requirements for deployment (see Define Release Criteria).		
	Implement the organization's staged deployment plan (see Plan Deployment), making adjustments based on testing results as necessary.		
	Adopt a continuous integration strategy that uses automated builds and deployment to facilitate frequent system updates [961].		
	Employ secure automated deployment methods, such as through an infrastructure as code approach [962].		
	Follow established guidelines to ensure the secure deployment of AI systems [963].		

Engage

The Engage stage represents the transition from the deployment phase into operations. Important throughout the adoption life cycle, this stage focuses on the organization's engagement with internal and external stakeholders. By engaging in these collaborative efforts, the organization helps to support a safer and more trustworthy AI ecosystem, build trust with its stakeholders, and identify potential issues when they arise.

	Engage	Owner	Guidance
Organization	**Government**	Leadership	SI.5, SI.7
	Support international collaboration efforts on AI risk management and governance [964, 965].		
	Ensure responsible use of the organization's AI systems if deployed for military applications [966, 967, 968].		
	Contribute to public AI information sharing [969], transparency and accountability [970], and incident disclosure [971, 972] initiatives.		
	Society	Responsibility	SI.1, SI.3
	Determine whom within civil society and the public the organization should engage with on the responsible use of AI systems [973].		
	Assess the range of public engagement mechanisms and select those best suited to reach the intended stakeholders [974].		
	Commit to responsible AI principles or other similar guidance and publicly publish those commitments [975].		
	Develop technology designed to promote rather than subvert social and civic processes [976].		
	Academia	Innovation, Cybersecurity	SI.1, SI.4, SI.5, VN.1
	Assess the opportunities and challenges in collaborating with academic institutions on AI research [459].		
	Enable structured access to AI models for researchers to help further the state of AI safety and security research [977].		
	Establish bug bounty programs to incentivize researchers to identify potential vulnerabilities or safety issues [978].		

	Engage	Owner	Guidance
Operations	**Stakeholders**	Communication, Marketing	SI.4, RC.2, ST.1, ST.3, ST.4, PP.10
	Maintain awareness of the sentiment and opinion of the general public on issues pertaining to AI technology and use [979, 980].		
	Engage with and gather input from a wide range of diverse stakeholders throughout the AI adoption life cycle [981].		
	Include secondary stakeholders, such as data contributors and affected communities, in engagement activities [982].		
	Work with stakeholders to proactively address concerns and provide mechanisms for transparency and redress [983].		
	Manage the expectations that stakeholders have related to AI adoption and provide mechanisms to collect feedback [984].		
Workforce	**Personnel**	Workforce	RC.4, IR.3, VN.1
	Include personnel in the process of AI adoption and provide opportunities for them to share and deliberate on how AI is used [985].		
	Make confidential issue reporting [986] and whistleblowing [987] mechanisms readily and meaningfully available to employees.		
	Help personnel develop basic generative AI competencies, including how to appropriately use AI and identify synthetic media [988].		
	Support personnel in adopting AI tools and work to strategically integrate AI initiatives in ways that advance their work [989].		
Infrastructure	**Service Providers**	Supply Chain	IM.2
	Work with service providers to mitigate security risks to the organization and collaborate on security incidents [990].		
	Engage service providers in the development of organizational AI supply chain risk management models [991].		
	Map out dependency networks (see Map Supply Chain) and engage service providers in conversations around accountability [992].		
Data	**Data Subjects**	Privacy, Data	PI.4, PI.5, PP.10
	Be transparent and public about how users' data is collected and used (see Communicate Data Practices).		
	Create channels for data subjects to request more information and lodge grievances [993, 994].		
	Provide mechanisms for individuals to object to [995] and opt out [996] of the collection and processing of their personal data.		
	Incorporate contestability into AI and algorithmic design [997]. Provide mechanisms to contest automated decision-making [998].		

	Engage	Owner	Guidance
Model	**Collaborative Initiatives**	Innovation, Cybersecurity	SI.4, SI.7, MO.5
	Participate in information sharing (see Monitor Information Sharing) and incident reporting (see Respond Communication) initiatives.		
	Contribute to collaborative AI governance and standards-setting programs [999].		
	Consider participating in regional or sector-specific cybersecurity mutual assistance programs, where available [1000, 1001].		
	Consider federated learning collaborations that enable institutions to jointly train systems while preserving privacy [1002].		
System	**End Users**	Product, Communication	SI.2
	Assess the factors influencing AI acceptance and take steps to address potential hurdles in adoption by users [878].		
	Provide end users with information on how to effectively use the AI system and training on its appropriate application [1003].		
	Establish clear mechanisms for users to provide feedback, report issues, and engage in auditing the AI system [1004, 1005].		
	Work with end users to calibrate an appropriate level of trust in the output of AI systems and combat automation bias [1006, 1007].		

Manage

The Manage stage represents the day-to-day operations of the organization. This includes activities to support the continued function, safety, and security of the organization's infrastructure and systems. This stage coincides with the monitoring efforts described in the following stage, as both of these operations represent ongoing activities.

	Manage	Owner	Guidance
Organization	**Situational Awareness**	Leadership	RC.1, MG.5, PM.5, MO.5
	Identify the information needs of the organization's leadership in order to maintain awareness of the organization's operations [1008].		
	Prioritize the sources of information needed for situational awareness, including threat intelligence and vulnerability assessment [1009].		
	Implement a program for collecting, synthesizing, and distributing actionable information (see Develop Management Systems) [1010].		
Operations	**Relationships**	Supply Chain	SC.1, RC.3
	Build relationships with stakeholders (see Engage Stakeholders) and develop relationship management practices [1011, 1012].		
	Implement supply chain management practices to cultivate and maintain relationships with suppliers and customers [1013].		
	Communicate information about the organization's adoption and use of AI systems to both internal and external stakeholders [1014].		
Workforce	**Personnel Awareness**	Workforce, Cybersecurity, Trust & Safety	WF.1, SC.4, FS.5, TO.1
	Provide cybersecurity training and awareness to all personnel within the organization [1015].		
	Disclose to employees what information is collected about them and their activity as part of enterprise security practices [1016].		
	Work with personnel to develop their AI literacy, including awareness of AI capabilities and limitations [1017, 1018, 1019].		
	Ensure that personnel are aware of the risks of synthetic media and its use in social engineering and misinformation [1020, 1021].		
	Provide situational awareness and information sharing mechanisms to support personnel working in human-machine teams [1022].		
Infrastructure	**Assets**	Infrastructure	NS.1, IV.4, SM.4, ES.1, IT.10
	Keep the asset inventory (see Map Assets) up-to-date and ensure all AI-related assets are accurately recorded [1023, 1024].		
	Schedule regular assurance checks and audits of the organization's AI inventory [527].		
	Employ asset management best practices (see Develop Management Systems) to maintain IT assets over their lifetime.		

	Manage	Owner	Guidance
Infrastructure	**Access**	Cybersecurity	SM.6, AC.3, AC.4, AC.5, AC.6, AC.8, AC.9, IA.2, IA.8
	Implement mechanisms to encourage regular review of user access and access updates when needs or roles change [1025].		
	Maintain access logs in accordance with organizational policies and legal requirements (see Monitor Access) [1026].		
	Deconflict access management policies made or implemented by multiple teams or stakeholders [1027].		
	Maintenance	Infrastructure	IM.8, ES.7, ES.8
	Keep a regular maintenance schedule in accordance with a severity or prioritization system [1028].		
	Consider using predictive maintenance to enhance the resilience and availability of systems [1029].		
	Consider maintenance approaches that balance security, risk management, cost, and efficiency [1030].		
	Decommission	Infrastructure	NS.8, MG.6, MS.1, PI.2
	Decommission assets and revoke permissions once they are no longer needed [524].		
	Employ a consistent process for decommissioning assets that accounts for dependencies and stakeholder communication [1031].		
	Update inventories once assets are decommissioned and monitor for unforeseen impacts resulting from decommissioning [1032].		
Data	**User Control**	Privacy, Data	PI.4, PI.5
	Provide mechanisms for users to view data that the organization has collected about them and how it has been processed [1033].		
	Respond to data subject access requests from users, or entities operating on their behalf, in a timely manner [1034].		
	Honor and comply with user requests to modify or delete personal data [1035]. Respect an individual's right to be forgotten [1036].		
	Comply with data control requirements (see Assess Legality) of the jurisdictions in which personal data is collected or processed [1037].		
Model	**Performance**	Product	PM.5, IT.10
	Review model performance (see Monitor Performance) and assess whether recalibrating, refitting, or retraining is needed [1038].		
	Use techniques such as unlearning [1039], fine-tuning [1040], or model editing [1041] to correct and align model behavior.		
	Consider techniques and strategies to automate model corrections [1042].		
	Detect and manage issues of data and model drift that can reduce the performance of model inference [1043].		

	Manage	Owner	Guidance
System	**Vulnerabilities**	Cybersecurity, Product	VN.3, VN.4, VN.5, VN.6
	Implement a standard process managing vulnerabilities in organizational and third-party assets [1044].		
	Maintain awareness of known risks and vulnerabilities [1045], including those specifically related to AI systems [47, 1046, 1047].		
	Identify vulnerabilities via scanning [1048], testing (see Test Security), and information sharing (see Monitor Information Sharing).		
	Investigate, triage, and remediate detected vulnerabilities in a systematic way, prioritizing the most critical first [1049, 1050].		
	Mitigate the risk of vulnerabilities that have not been, or cannot, be patched [1051]. AI systems inherently present these risks [1052].		
	Communicate and share information about vulnerabilities, including those found in AI systems [1053].		

Monitor

The Monitor stage covers the range of ongoing activities intended to detect issues as they arise. These issues include a range of safety, security, privacy, compliance, and performance problems. These activities, like those of the previous Manage stage, make up the ongoing operations of the organization. When an issue is detected, that triggers the Respond stage, which is described in the next section.

	Monitor	Owner	Guidance
Organization	**Context**	Legal	AU.2
	Monitor the evolving legal and regulatory landscape as it pertains to AI to identify changes that impact the organization [1054, 1055].		
	Understand the legal and regulatory context governing which activities or events can be monitored and logged [546].		
	Establish an authority who can respond to requests for information from law enforcement or other external stakeholders [1056].		
	Physical Environment	Physical Security	PS.6, PS.3, MO.6, SM.6
	Monitor physical access to the organization's facilities and centralize physical security functions within an internal security office [587].		
	Coordinate regular security risk assessments to assess priorities and avoid duplication of efforts [1057].		
	Monitor the physical environment to detect and prevent environmental hazards and safety incidents [1058].		
Operations	**Centralized Analysis**	Operations, Cybersecurity	MO.1, MO.3, MO.7, IR.8, LG.3, VN.4
	Establish a security information and event management platform for centralized event and log analysis [1059].		
	Prioritize what information and logs are included in the centralized analysis repository [1060].		
	Conduct ongoing vulnerability and security log analysis to detect potential security events [1061].		
	Enhance analysis capabilities by using predictive analytics and AI/ML techniques [1062].		

	Monitor	Owner	Guidance
Operations	**Information Sharing**	Cybersecurity	MO.2, MO.5, IR.6, VN.1, VN.4, IR.3, IR.6, ES.5, IS.5
	Collect, analyze, and share cyber threat intelligence to inform security decisions and maintain a proactive security posture [1063, 1064].		
	Participate in cybersecurity information sharing initiatives to receive and provide rapid information about emerging attacks [1065].		
	Consider joining the organization's sector-specific information sharing and analysis center or organization [1066].		
	Report incidents and near misses through established incident reporting channels [1067].		
	Third Parties	Supply Chain, Cybersecurity	AU.5, SC.7, PS.2
	Monitor third-party access [1068] to the organization's systems and facilities and their activity while having access [1069].		
	Monitor suppliers to ensure their continued compliance and that they uphold service level agreements [1070, 1071].		
	Establish mechanisms to share pertinent information with third-party suppliers [1072, 1073].		
	Use predictive analytics to improve the monitoring and identification of supply chain threats [1074].		
Workforce	**Security Operations**	Cybersecurity	MO.1, MO.8
	Establish a security operations center or similar team to monitor and analyze potential incidents across the organization [1075].		
	Identify and employ personnel with institutional knowledge and subject matter expertise in various monitoring domains [1076].		
	Ensure that personnel with monitoring responsibilities are placed in the correct teams and under the proper reporting structure [1077].		
	Insider Threats	Physical Security, Cybersecurity	MO.6, AC.9, LG.4
	Understand the indicators of insider threats and establish mechanisms to systematically measure them [1078].		
	Monitor personnel activity and access patterns to identify potential insider threats [1079, 1080, 1081].		
Infrastructure	**Access**	Cybersecurity	MO.4, LG.2, SM.6, IA.6, AC.9
	Validate access patterns against established access control rules to identify violations and anomalies (see Control Access).		
	Monitor users' interactions with the organization's systems and data through access control logs and similar audit mechanisms [1082].		
	Perform periodic reviews of access control policies. Update account permissions in accordance with least privilege principles [1083].		

	Monitor	Owner	Guidance
Infrastructure	**Data Flows**	Infrastructure, Cybersecurity	MO.4, IS.5, LG.2
	Monitor the organization's networks and log events that capture cybersecurity-relevant information for analysis [547, 1084].		
	Employ anomaly detection techniques to identify unusual and suspicious activity on the organization's networks [1085].		
	Set appropriate thresholds for anomaly detection and recalibrate regularly [1086]. Consider using adaptive dynamic thresholds [1087].		
	Monitor for indicators of network compromise [1088]. Update the set of known indicators in a regular or automated fashion [1089].		
Data	**Data Quality**	Data	TR.5
	Determine appropriate indicators of data quality and employ metrics to measure them over time [558, 1090, 856].		
	Implement a data management pipeline that enables continuous monitoring to assess and detect data quality issues [1091].		
	Privacy	Privacy, Data	PI.7, PI.8
	Monitor network traffic, transfers, and access logs to detect potential data breaches or unintended data exposure [1092].		
	Employ continuous privacy monitoring at the organization, business-process, and information-system level [1093].		
	Ensure that monitoring and logging comply with privacy and data retention best practices [1094].		
Model	**Performance**	Product	PM.1, PM.2, TO.4, SG.4
	Select metrics (see Define Metrics) and appropriate benchmarks to track model performance over time [1095].		
	Design a continuous monitoring strategy that tracks outcomes across different subgroups and environments [1096].		
	Employ automated methods of tracking and identifying model drift [1097].		
	Monitor for performance anomalies and use collected data to identify bottlenecks in system processing [1098].		
	Develop a post-deployment performance baseline and monitor for substantial deviations [1099].		
	Alignment	Product	PM.2
	Log and monitor AI model alignment and set thresholds for misalignment where appropriate and feasible [1100].		
	Adopt or create scoring metrics for model alignment that are tailored to the organization's specific use case [1101].		
	Consider integrating self-monitoring mechanisms into a model's chain-of-thought reasoning process [1102].		

	Monitor	Owner	Guidance
Model	**Fairness**	Trust & Safety	FS.2, FS.4, TO.4
	Establish metrics to continuously monitor issues of bias and fairness [1103].		
	Implement run-time monitoring for certain fairness metrics that may differ between development and deployment contexts [1104].		
	Consider how different methods for evaluating fairness may lead to different outcomes or may be inappropriate in given contexts [1105].		
	Inputs & Outputs	Product	SG.3, IS.4
	Monitor model inputs and outputs for unanticipated issues or bypassed safeguards (see Control Inputs and Control Outputs).		
	Monitor the relationship between inputs and outputs to detect malicious attacks on AI models [1106].		
	Account for harm [1107] and malicious use [1108] that can occur across multiple interactions or be distributed across multiple accounts.		
System	**Behavior**	Product	PM.2, SG.4, SG.6, LG.5
	Establish a baseline of the system's normal operating behavior to aid in the identification of anomalous behavior [1109].		
	Review and update the baseline for system behavior and the safety and security thresholds used to trigger a response [1110].		
	Consider using adaptive thresholds for anomaly detection in complex systems and environments [1111, 1112].		
	Implement methods to facilitate real-time failure detection for systems that are operating with some level of autonomy [1113].		
	Use	Trust & Safety	AU.5, TO.4, LG.5
	Establish criteria for acceptable system use or indicators of misuse. Monitor these metrics to flag potential misuse [1114].		
	Consider the use of user and entity behavior analytics to detect anomalous user behavior [1115].		
	Identify cases when the AI system is used for purposes other than what it was intended for that may lead to harm [1116].		
	Impact	Trust & Safety	IM.2, IM.8
	Identify relevant impact metrics and design mechanisms to monitor them over time [444].		
	Measure negative downstream impacts of AI systems including societal [1117], environmental [1118], and health [1119] impacts.		
	Account for attritional harms from AI that develop gradually over time and can be harder to identify and measure [1120].		
	Measure and promote positive downstream impacts of AI [1121].		

Respond

The Respond stage is unique, as its occurrence is contingent on a potential issue being detected during the course of the organization's ongoing monitoring operations. The guidance in this stage reflects general incident response practices and the nuances in responding to issues involving or stemming from AI systems.

	Respond	Owner	Guidance
Organization	**Remediation**	Legal	RC.4, RC.6, TO.7
	Identify the individuals and stakeholders that were impacted as a result of the indecent, adverse event, or other harm [1122].		
	Provide for or cooperate in remediation efforts and comply with any judicial remediation decisions [1123, 1124].		
	Ensure that the organization provides meaningful redress for AI harms and works to translate those efforts into systemic change [1125].		
Operations	**Triage**	Operations	IR.3, IR.4
	Develop an incident severity scale, rubric, or classification framework to aid in the triage of potential incidents [1126, 1127].		
	Map alerting and monitoring systems to the incident severity scale (see Monitor Centralized Analysis).		
	Employ automated tools to support human security analysts in triaging the often overwhelming number of alerts [1128, 1129].		
	Provide decision support to aid response teams in balancing individual event triage and broader operational awareness [1130].		
	Communication	Communication	IR.1, IR.6, MO.5, IR.7, RR.3, PS.5, PI.8
	Support crisis management activities by communicating incident information, status, and strategic decisions to stakeholders [1131].		
	Designate responsibilities and coordinate response activities among all parties involved in responding to the incident [1132].		
	Recognize that insider events can present additional challenges that require a specialized approach to communications [1133].		
	Report cybersecurity incidents in accordance with legal requirements [1134]. Report AI incidents as channels are created [1135, 1136].		
	Communicate information about the incident and its response to stakeholders post-incident and work to repair public relations [1137].		
	Document information about the incident, response activities, and evidence collected during the investigation [1138].		

	Respond	Owner	Guidance
Workforce	**Response Team**	Workforce	IR.1, IR.2, MO.8
	Define a core response team [1139] and a broader set of personnel that can be pulled in to provide specific expertise [1140].		
	Ensure the broader response team includes personnel with AI expertise and upskill the entire team on AI-related incidents [1141].		
	Foster adaption, collective problem-solving, communication, trust, and shared knowledge among response team members [1142].		
	Provide adequate guidance and empower response team members to make critical decisions related to incident response [1143].		
	Support the incident response team and take steps to address issues of burnout that can be high among response staff [1144].		
	Alerting	Communication	MO.1, MO.8, IR.2
	Establish a rotation of personnel for incident monitoring and on-call staff to alert in case of emergencies [1145].		
	Create incentives for employees to report anomalies, alerts, or suspicious events. Do not unnecessarily punish false alarms [1146].		
	Establish multiple communications channels to rapidly alert relevant personnel of incidents when they arise [1147].		
Infrastructure	**Containment & Neutralization**	Operations, Infrastructure	IR.4, RR.9, MO.3, MO.6, IR.2
	Employ segmented and compartmentalized security domains to quickly lock down systems and contain security threats [1148].		
	Consider employing automated and AI-based incident detection and response mechanisms for more rapid containment [1149].		
	Respond to AI incidents quickly, applying deployment corrections to address dangerous capabilities, behavior, or misuse [1150].		
	Neutralize and remove the source of the incident once contained [1151].		
	Investigation	Operations, Cybersecurity	IR.5, IR.8, PI.8
	Initiate a computer forensics investigation to identify the cause of the incident, the scope of impact, and potential attribution [1152, 1153].		
	Identify aspects of AI systems that may require AI-specific digital forensic analysis [1154] or an agent-specific analysis [1155].		
	Consider the use of ML and AI technologies to augment digital forensics teams [1156].		
	Conduct backward analysis to identify if similar undetected issues occurred or similar vulnerabilities remain [1157].		
	Cooperate with government investigations when significant cyber incidents (e.g., widespread or critical outages) occur [1158].		

	Respond	Owner	Guidance
Data	**Data Breach**	Privacy, Cybersecurity	PI.8
	Follow response checklists [1159] when responding to data breaches, including specific guides for ransomware [1160] incidents.		
	Comply with legal obligations regarding notification about data breaches involving personal data [1161].		
	Investigate cases where data breaches involve AI systems, including where shadow AI use by employees caused the breach [1162].		
	Restoration	Data	RR.9
	Validate backup integrity and ensure recoverability before attempting to restore data [1163, 1164].		
	Use forensic recovery tools and techniques in cases where backups are missing, partially corrupted, or insufficient [1165, 1166].		
Model	**Fail-Safes**	Product	SG.1
	Develop fail-safes such that the AI system is able to revert to safe states when encountering various failure modes [1167].		
	Employ human-in-the-loop mechanisms for AI systems [1168] and defer to human oversight when issues or failures occur [1169].		
	Understand the conditions where various forms of human oversight and control over AI is possible and where it is limited [1170].		
	Restoration	Product	RR.9, SG.3, SG.6, PM.4
	Investigate and address issues of corruption [1171] or compromise [1172, 1173] in models and backups before restoration.		
	Implement pipelines that capture snapshots of model states to facilitate rapid restoration after an incident occurs [1174].		
System	**Resilience Mechanisms**	Product	RR.6, SG.1
	Implement checkpoints to save operational state and mechanisms to roll back to that state in the case of failures [1175].		
	Consider employing hardware, software, data, network, infrastructure, and utility redundancy [1176].		
	Employ failover, load balancing, and redundancy to ensure the high availability of critical systems [1177].		
	Ensure that the system can be safely and promptly shut down when needed, particularly if AI based [1178].		
	Restoration	Product	RR.9
	Perform walk-throughs of system and data restoration procedures with key employees and stakeholders pre-incident [1179].		
	Triage system functions and prioritize restoring the most important functionality first, especially in safety-critical contexts [1180].		
	Verify the integrity of systems and ensure that they are restored to a safe state before resuming operations [1181].		

Improve

The Improve stage represents the transition from normal operations back to the review phase in which the organization started. This stage should be triggered at defined intervals, after an incident occurs and the related response completes, and before the organization sets out to adopt additional or more advanced AI systems. At this stage, the organization should review its practices to identify and implement improvements—incorporating lessons learned from the previous adoption cycle and from any incidents that have arisen. Improvements that require major changes to the AI system may warrant starting the cycle from the self-assessment stage once more.

	Improve	Owner	Guidance
Organization	**Strategy**	Leadership	MG.4, MG.8, SL.3, SM.1, TO.5, AU.2
	Conduct an honest and open business strategy review to determine whether and how the strategy should change [34].		
	Identify gaps in knowledge and awareness that can be addressed across the organization in order to better inform strategy [1182].		
	Prioritize an organizational strategy that integrates crisis management [1183] and builds resilience against future incidents [1184].		
Operations	**Procedures**	Operations	MG.4, MG.8, Sl.3, RM.9, AU.4
	Conduct regular audits of the management system and IT processes to identify improvements [1185].		
	Collaborate with personnel to generate process improvement ideas [1186].		
	Assess whether resource utilization and allocation are appropriate and whether improvements can be made [53].		
	Learn strategically from incidents and near misses, incorporating lessons learned to update policies and procedures [1187].		
	Monitoring	Operations	IM.8, MO.4, MO.5
	Conduct a post-incident review to refine monitoring systems and evaluate whether existing processes are sufficient [1188].		
	Update organizational knowledge and personnel awareness after incidents occur to improve future detection [1189].		
	Response	Operations	PM.4, IR.5, IR.9, RR.2
	Review past incidents and response activities to identify lessons learned [1190].		
	Update the incident response plan (see Plan Response) based on the lessons learned from previous incidents [1191].		

	Improve	Owner	Guidance
Operations	**Supply Chain**	Supply Chain	SC.6, SC.7
	Establish metrics to track the performance of the organization's supply chain management [1192].		
	Conduct regular audits of the supply chain and supplier practices [1193].		
	Engage employees [1194] and customers [1195] to identify improvements along the organization's supply chain.		
Workforce	**Training & Awareness**	Workforce	WF.1, ST.5, SC.4
	Review and evaluate the effectiveness of the organization's performance management system [1196].		
	Evaluate the effectiveness of training programs [1197] and cybersecurity awareness initiatives [1198].		
	Reassess talent gaps (see Define Needs) and adapt training programs to address them, as necessary [1199].		
Infrastructure	**Controls**	Cybersecurity	IS.6, IR.9, RM.8, ST.3
	Identify root causes of incidents and update controls to prevent or mitigate future occurrences [1200, 1201].		
	Align control improvement plans with threat intelligence, frameworks, and other resources that impact organizational strategy [1202].		
	Collect metrics for access control performance and use them to improve access control across the organization's systems [1203].		
	Conduct regular audits of the organization's cybersecurity program, controls, and protections [1204, 1205].		
Data	**Data Handling**	Data, Privacy	PP.2
	Conduct data audits on a regular basis to review the organization's data management practices [1206].		
	Improve data governance capabilities to enhance the development of future systems and promote trustworthiness [1207].		
	Investigate cases where data has been mishandled or accidentally exposed and work to ensure that it does not occur again [1208, 1209].		
Model	**Outcomes**	Product, Cybersecurity	PM.5, TO.5
	Assess and correct the alignment of the AI model with human values and organizational goals on a regular basis [1210].		
	Review the decision-making and improve the accountability of AI systems [1211] and human-AI teams [1212].		
	Update model guardrails and broader risk controls (see Control Risk) after an incident to prevent similar issues in the future [1213].		

	Improve	Owner	Guidance
Model	**Testing**	Innovation	SG.4, TE.5, PM.3, IM.8
	Stay up-to-date with test and evaluation best practices, as these continue to evolve for AI systems [1214].		
	Continue to conduct testing of the AI model and system throughout its life cycle, not just at the end of model development [1215].		
	Review and improve on the technical, procedural, and organizational aspects of software system testing [1216].		
System	**System Design**	Product, Innovation	PM.5
	Analyze system failures to identify improvements and inform the requirements and design of future systems [1217].		
	Conduct a postmortem analysis of the system once in production to identify improvement for future projects [1218].		
	Apply a continuous improvement approach to the organization's software and systems [1219].		
	User Experience	Product	IR.9, SI.2
	Incorporate feedback from users (see Engage End Users) on their experience to identify system improvements [1220, 1221].		
	Assess usability [1222] and user experience [1223] to identify improvements to the AI system [1224].		
	Promote widespread and equitable access to the organization's AI tools and services [1225].		

Conclusion

In this report, we provide an extensive reference guide to support practitioners in operationalizing AI best practices within their organizations. In doing so, we help to answer the core questions required for implementation: who (the business function responsible), what (the implementation step with its detailed recommendations), when (at which stage in the adoption life cycle), where (at what implementation level within the organization), why (which recommendation from the harmonized framework does it implement), and how (the external resources). Altogether, this guide identifies 238 implementation steps, 817 detailed recommendations, and 1,225 external resources to support operationalization. As a whole, this guide represents a comprehensive picture of AI adoption, but for many practitioners only a subset of this information will be immediately relevant. As such, we present this guide in a structured manner to enable practitioners to easily identify implementation steps that are pertinent to their use case, and we provide mechanisms for quick navigation across the report.

This report represents the second step in addressing the challenges organizations face in implementing AI guidance. Building on CSET's initial research that harmonized AI guidance into a single unified framework, this work provides practical steps to implement that harmonized guidance [10]. In doing so, we help to address the lack of implementation details that are common in AI guidance documents and frameworks. To aid practitioners further, future research is needed to tailor these recommended practices to the unique aspects of various AI use cases, sectors, and types of organizations.

Authors

Kyle Crichton is a research fellow at CSET, where he works on the CyberAI Project focusing on security and privacy challenges related to AI systems.

Abhiram Reddy completed his contributions to this research while he was a student research assistant at CSET.

Jessica Ji is a senior research analyst at CSET, where she works on the CyberAI Project focusing on AI red teaming and AI governance.

Acknowledgments

Thank you to Drew Lohn and John Bansemer for their support and guidance in conducting this research. We would also like to thank Catherine Aiken, Matt Mahoney, Nikhil Mulani, and Jonathan Spring for their helpful feedback during the review process. This research was supported in part by generous funding from the AI Safety Fund and a Google Academic Research Award (GARA).

Document Identifier: doi: 10.51593/20250002

References

[1] Teemu Birkstedt, Matti Minkkinen, Anushree Tandon, and Matti Mäntymäki, "AI Governance: Themes, Knowledge Gaps and Future Agendas," *Internet Research* 33, no. 7 (18 December 2023): 133–167, https://doi.org/10.1108/INTR-01-2022-0042.

[2] Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon, "Principles to Practices for Responsible AI: Closing the Gap," arXiv preprint arXiv:2006.04707 (2020), https://doi.org/10.48550/arXiv.2006.04707.

[3] Janek Bevendorff, Matti Wiegmann, Martin Potthast, and Benno Stein, "Is Google Getting Worse? A Longitudinal Investigation of SEO Spam in Search Engines," in *Advances in Information Retrieval: 46th European Conference on Information Retrieval*, eds. Nazli Goharian, Nicola Tonellotto, Yulan He et al. (Springer-Verlag, 2024): 56–71, https://doi.org/10.1007/978-3-031-56063-7_4.

[4] Hugh Langley, "Google Recently Cut 'People' from Its Search Guidelines. Now, Website Owners Say a Flood of AI Content Is Pushing Them Down in Search Results," *Business Insider*, September 20, 2023, www.businessinsider.com/google-search-helpful-content-update-results-drop-ai-generated-2023-9.

[5] Kate Knibbs, "AI Slop Is Flooding Medium," *Wired*, October 28, 2024, www.wired.com/story/ai-generated-medium-posts-content-moderation/.

[6] National Institute of Standards and Technology, *AI RMF Playbook* (U.S. Department of Commerce, 2024), https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook.

[7] OWASP Foundation, "OWASP AI Exchange," last updated July 4, 2025, accessed August 2025, https://owaspai.org/.

[8] OWASP Foundation, "OWASP AI Testing Guide," accessed October 2025, https://owasp.org/www-project-ai-testing-guide/.

[9] Stephanie Ifayemi, Elham Tabassi, and Amanda Craig Deckard, "Decoding AI Governance: A Toolkit for Navigating Evolving Norms, Standards, and Rules," Partnership on AI and Microsoft, November 19, 2024, https://partnershiponai.org/resource/decoding-ai-governance/.

[10] Kyle Crichton, Abhiram Reddy, Jessica Ji et al., *Harmonizing AI Guidance: Distilling Voluntary Standards and Best Practices into a Unified Framework* (CSET, July 2025), https://doi.org/10.51593/20240041.

[11] Shahar Avin, Miles Brundage, Gretchen Krueger et al., "Filling Gaps in Trustworthy Development of AI," *Science* 374, no. 6473 (December 2021): 1327–1329, https://doi.org/10.1126/science.abi7176.

[12] Jianlong Zhou and Fang Chen, "AI Ethics: From Principles to Practice," *AI & Society* 38 (2023): 2693–2703, https://doi.org/10.1007/s00146-022-01602-z.

[13] Malak Sadek, Emma Kallina, Thomas Bohné, Céline Mougenot, Rafael A. Calvo, and Stephen Cave, "Challenges of Responsible AI in Practice: Scoping Review and Recommended Actions," *AI & Society* 40 (2025): 199–215, https://doi.org/10.1007/s00146-024-01880-9.

[14] Lionel Nganyewou Tidjon and Foutse Khomh, "The Different Faces of AI Ethics Across the World: A Principle-to-Practice Gap Analysis," *IEEE Transactions on Artificial Intelligence* 4, no. 4 (August 2023): 820–839, https://doi.org/10.1109/TAI.2022.3225132.

[15] Steven Mills, Elias Baltassis, Maximiliano Santinelli, Cathy Carlisi, Sylvain Duranton, and Andrea Gallego, "Six Steps to Bridge the Responsible AI Gap," BCG, September 8, 2020, www.bcg.com/publications/2020/six-steps-for-socially-responsible-artificial-intelligence.

[16] Kyle Crichton, Jessica Ji, Kyle Miller et al., *Securing Critical Infrastructure in the Age of AI* (CSET, October 2024), https://cset.georgetown.edu/publication/securing-critical-infrastructure-in-the-age-of-ai/.

[17] Elizabeth A. Lynch, Alison Mudge, Sarah Knowles, Alison L. Kitson, Sarah C. Hunter, and Gill Harvey, "There Is Nothing So Practical as a Good Theory: A Pragmatic Guide for Selecting Theoretical Approaches for Implementation Projects," *BMC Health Services Research* 18, no. 857 (2018), https://doi.org/10.1186/s12913-018-3671-z.

[18] Enola K. Proctor, Nyron J. Powell, and J. Curtis McMillen, "Implementation Strategies: Recommendations for Specifying and Reporting," *Implementation Science* 8, no. 139 (2013), https://doi.org/10.1186/1748-5908-8-139.

[19] Wajid Ali and Abdul Zahid Khan, "Factors Influencing Readiness for Artificial Intelligence: A Systematic Literature Review," *Data Science and Management* 8, no. 2 (2024): 224–236, https://doi.org/10.1016/j.dsm.2024.09.005.

[20] IT Modernization Center of Excellence, "AI Capability Maturity," in *AI Guide for Government* (U.S. General Services Administration, 2025), ch. 6, https://coe.gsa.gov/coe/ai-guide-for-government/ai-capability-maturity/index.html.

[21] James Baney, John Mahoney, Kevin Inks, and Yee San Su, *CNA's Artificial Intelligence (AI) Maturity Model for Government Agencies* (CNA, May 2025), www.cna.org/analyses/2025/05/artificial-intelligence-maturity-model.

[22] Matteo Meucci, Philippe Schrettenbrunner, Arvinda Gangadhararao et al., *OWASP AI Maturity Assessment* (OWASP Foundation, August 2025), https://owasp.org/www-project-ai-maturity-assessment/.

[23] Eric E. Bloedorn, Diane M. Kotras, Peter J. Schwartz, Clyneice Chaney, Clareice Chaney, and Jim Patsis, *The MITRE AI Maturity Model and Organizational Assessment Tool Guide: A Path to Successful AI Adoption* (MITRE Corporation, November 2023), https://aimaturitymodel.mitre.org/.

[24] Chief Digital and Artificial Intelligence Office, "RAI Toolkit," U.S. Department of War, accessed August 2025, www.tradewindai.com/rai-toolkit.

[25] Commission Nationale de l'informatique et des Libertés, *Self-Assessment Guide for Artificial Intelligence (AI) Systems* (CNIL, August 2022), www.cnil.fr/en/self-assessment-guide-artificial-intelligence-ai-systems.

[26] Cisco, "AI Readiness Index," accessed August 2025, www.cisco.com/c/m/en_us/solutions/ai/readiness-index/assessment-tool.html.

[27] VTT Technical Research Centre of Finland, "Artificial Intelligence (AI) Maturity Tool," accessed February 2026, https://ai.eitcommunity.eu/services/ai-maturity-tool.

[28] Fifth Quadrant, "Responsible AI Index," accessed August 2025, www.fifthquadrant.com.au/responsible-ai-index.

[29] Raj Vayyavur, "Why AI Projects Fail: The Importance of Strategic Alignment and Systematic Prioritization," *International Journal of Research* 11 (2025): 386–391, https://doi.org/10.5281/ZENODO.13370566.

[30] John C. Henderson and N. Venkatraman, *Strategic Alignment: A Process Model for Integrating Information Technology and Business Strategies* (Center for Information Systems Research, Massachusetts Institute of Technology, 1989), https://dspace.mit.edu/bitstream/handle/1721.1/49087/strategicalignmex00hend.pdf.

[31] Jonathan H. Reed, "Modeling and Measuring Strategic Alignment," *Journal of Strategy and Management* 16, no. 4 (November 2023): 654–671, https://doi.org/10.1108/JSMA-11-2022-0212.

[32] Tim McLaren, Milena Head, Yufei Yuan, and Yolande E. Chan, "A Multilevel Model for Measuring Fit Between a Firm's Competitive Strategies and Information Systems Capabilities," *MIS Quarterly* 35 (2011): 909–929, https://doi.org/10.2307/41409966.

[33] Martin Smits, Alea Fairchild, Pieter Ribbers, Koen Milis, and Erik van Geel, "Assessing Strategic Alignment to Improve IT Effectiveness," *BLED Proceedings* (2009), https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1027&context=bled2009.

[34] Russell A. Eisenstat and Michael Beer, "How to Have an Honest Conversation About Your Business Strategy," *Harvard Business Review* 82, no. 2 (February 2004): 82–89, https://hbr.org/2004/02/how-to-have-an-honest-conversation-about-your-business-strategy.

[35] Pitabas Mohanty, Supriti Mishra, and Tina Stephen, "AI Governance Maturity Matrix: A Roadmap for Smarter Boards," *California Management Review* 67, no. 2 (2025), https://cmr.berkeley.edu/2025/05/ai-governance-maturity-matrix-a-roadmap-for-smarter-boards/.

[36] Ravit Dotan, Borhane Blili-Hamelin, Ravi Madhavan, Jeanna Matthews, Joshua Scarpino, Carol Anderson, *A Flexible Maturity Model for AI Governance Based on the NIST AI Risk Management Framework* (IEEE, July 2024), https://ieeeusa.org/product/a-flexible-maturity-model-for-ai-governance/.

[37] Monika Viktorova and Hadassah Drukarch, *Operationalizing Independent Review in AI Governance: A Guide for Practitioners* (Responsible Artificial Intelligence Institute, 2024), www.responsible.ai/operationalizing-independent-review-in-ai-governance/.

[38] Innovation Adoption Practice, "Innovation Adoption Culture Pre-assessment Guide," IT Modernization Centers of Excellence, May 24, 2022, https://coe.gsa.gov/2022/05/24/ia-update-2.html.

[39] Dirk Coetzee, "The Artificial Intelligence Readiness Prism: A Multi-dimensional Framework for Assessing AI Integration, Ethics, and Cultural Alignment," 2025, https://doi.org/10.13140/RG.2.2.30096.32003.

[40] Maximilian Röglinger, Jens Pöppelbuß, and Jörg Becker, "Maturity Models in Business Process Management," *Business Process Management Journal* 18, no. 2 (April 2, 2012): 328–346, https://doi.org/10.1108/14637151211225225.

[41] Patrick Mikalef and Manjul Gupta, "Artificial Intelligence Capability: Conceptualization, Measurement Calibration, and Empirical Study on Its Impact on Organizational Creativity and Firm Performance," *Information & Management* 58, no. 3 (2021), https://doi.org/10.1016/j.im.2021.103434.

[42] Liz Dyrsmid, "How to Evaluate Standard Operating Procedures for Improved Efficiency," Flowster, March 13, 2024, https://flowster.app/how-to-evaluate-standard-operating-procedures/.

[43] Azalia Shamsaei, Daniel Amyot, and Alireza Pourshahid, "A Systematic Review of Compliance Measurement Based on Goals and Indicators," in *Advanced Information Systems Engineering Workshops*, eds. Camille Salinesi and Oscar Pastor (Springer, 2011), https://doi.org/10.1007/978-3-642-22056-2_25.

[44] Lynda M. Bourne, "SRMM Stakeholder Relationship Management Maturity," paper presented at PMI Global Congress, St. Julians, Malta, May 19–21, 2008, www.stakeholdermapping.com/srmm-maturity-model/srmm-implementation/.

[45] Mohsen Cheshmberah and Safoura Beheshtikia, "Supply Chain Management Maturity: An All-Encompassing Literature Review on Models, Dimensions and Approaches," *Logforum* 16, no. 1 (2020): 103–116, https://doi.org/10.17270/J.LOG.2020.377.

[46] Oumaima Hansali, Samah Elrhanimi, and Laila El Abbadi, "Supply Chain Maturity Models: A Comparative Review," *Logforum* 18, no. 4 (2022): 435–450, https://doi.org/10.17270/J.LOG.2022.751.

[47] "Welcome to the AI Incident Database," AI Incident Database, accessed October 2025, https://incidentdatabase.ai/.

[48] "Voluntary AI Commitments," Biden White House Archives, September 2023, https://bidenwhitehouse.archives.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.

[49] Sangjae Lee and Hyunchul Ahn, “Assessment of Process Improvement from Organizational Change,” *Information and Management* 45, no. 5 (2008): 270–280, https://doi.org/10.1016/j.im.2003.12.016.

[50] Unanet, “Elevate Your Efficiency,” accessed October 2025, https://unanet.com/resource-management-maturity-model.

[51] Christopher P. Holland and Ben Light, “A Stage Maturity Model for Enterprise Resource Planning Systems Use,” *ACM SIGMIS Database* 32, no. 2 (Spring 2001): 34–45, https://doi.org/10.1145/506732.506737.

[52] Tom Butler, Angelina Espinoza-Limón, and Selja Seppälä, “Towards a Capability Assessment Model for the Comprehension and Adoption of AI in Organizations,” arXiv preprint arXiv:2305.15922 (May 25, 2023), https://doi.org/10.48550/arXiv.2305.15922.

[53] Susanne Hupfer, “Talent and Workforce Effects in the Age of AI,” Deloitte, March 3, 2020, www.deloitte.com/us/en/insights/topics/emerging-technologies/ai-adoption-in-the-workforce.html.

[54] National Association of State Chief Information Officers, *Enterprise Architecture Maturity Model* (NASCIO, December 2003), www.nascio.org/resource-center/resources/enterprise-architecture-maturity-model/.

[55] Nujud Alsufyani and Asif Qumer Gill, “A Review of Digital Maturity Models From Adaptive Enterprise Architecture Perspective: Digital by Design,” in *2021 IEEE 23rd Conference on Business Informatics* (IEEE, 2021), 121–130, https://doi.org/10.1109/CBI52690.2021.00023.

[56] Felix Klingenstein, Samuel Kessler, Moritz Hoeltl, Pascal Daume, and Marcus Fischer, “A Maturity Model to Assess and Enhance the AI Readiness of an Enterprise Architecture,” *AMCIS 2025 Proceedings* 27 (August 2025), https://aisel.aisnet.org/amcis2025/sig_osra/sig_osra/27.

[57] Edward Robertson, Gabrielle Peko, and David Sundaram, “Enterprise Architecture Maturity: A Crucial Link in Business and IT Alignment," *PACIS 2018 Proceedings* (2018), https://aisel.aisnet.org/pacis2018/308.

[58] U.S. Cybersecurity and Infrastructure Security Agency (CISA), “Zero Trust Maturity Model,” version 2.0, April 2023, www.cisa.gov/zero-trust-maturity-model.

[59] Dhavalkumar Patel, Ganesh Raut, Satya Narayan Cheetirala et al., “Cloud Platforms for Developing Generative AI Solutions: A Scoping Review of Tools and Services,” arXiv preprint arXiv:2412.06044 (2024), https://doi.org/10.48550/arXiv.2412.06044.

[60] Mark Paulk, Bill Curtis, Mary Chrissis, and C. V. Weber, “Capability Maturity Model,” *IEEE Software* 10 (July 1993): 18–27, https://doi.org/10.1109/52.219617.

[61] *ISO/IEC TS 33061:2021 Information Technology — Process Assessment — Process Assessment Model for Software Life Cycle Processes* (International Organization for Standardization, 2021), www.iso.org/standard/80362.html.

[62] Meenu Mary John, Helena Holmström Olsson, and Jan Bosch, "Towards MLOps: A Framework and Maturity Model," in *2021 47th Euromicro Conference on Software Engineering and Advanced Applications* (IEEE Computer Society, 2021): 1-8, https://doi.org/10.1109/SEAA53835.2021.00050.

[63] Mohammad Zarour, Hamza Alzabut, and Khalid T. Al-Sarayreh, "MLOps Best Practices, Challenges and Maturity Models: A Systematic Literature Review," *Information and Software Technology* 183 (2025), https://doi.org/10.1016/j.infsof.2025.107733.

[64] Seunghwan Cho, Ingyu Kim, Jinhan Kim, Honguk Woo, and Wanseon Shin, "A Maturity Model for Trustworthy AI Software Development," *Applied Sciences* 13, no. 8 (2023): 4771, https://doi.org/10.3390/app13084771.

[65] Rama Akkiraju, Vibha Sinha, Jalal Mahmud et al., "Characterizing Machine Learning Processes: A Maturity Framework," *Business Process Management* 12168 (2020), https://doi.org/10.1007/978-3-030-58666-9_2.

[66] James S. Pennypacker and Kevin P. Grant, "Project Management Maturity: An Industry-Wide Assessment," paper presented at PMI Research Conference 2002: Frontiers of Project Management Research and Applications, Seattle, Washington, www.pmi.org/learning/library/pm-maturity-industry-wide-assessment-9000.

[67] Office of Cybersecurity, Energy Security, and Emergency Response, *C2M2: Cybersecurity Capability Maturity Model*, version 2.1 (U.S. Department of Energy, June 2022), www.energy.gov/ceser/cybersecurity-capability-maturity-model-c2m2.

[68] Alenka Brezavšček and Alenka Baggia, "Recent Trends in Information and Cyber Security Maturity Assessment: A Systematic Literature Review," *Systems* 13, no. 1 (2025): 52, https://doi.org/10.3390/systems13010052.

[69] "Software Assurance Maturity Model," OWASP Foundation, accessed September 2025, https://owaspsamm.org.

[70] Manuel Kern, Max Landauer, Florian Skopik, and Edgar Weippl, "A Logging Maturity and Decision Model for the Selection of Intrusion Detection Cyber Security Solutions," *Computers & Security* 141 (2024), https://doi.org/10.1016/j.cose.2024.103844.

[71] American Institute of Certified Public Accountants and Canadian Institute of Chartered Accountants, *AICPA/CICA Privacy Maturity Model* (March 2011), https://iapp.org/media/pdf/resource_center/aicpa_cica_privacy_maturity_model_final-2011.pdf.

[72] Stuart Shapiro and Julie S. McEwan, "MITRE's Privacy Engineering Tools and Their Use in a Privacy Assessment Framework," news release, MITRE, November 13, 2019, www.mitre.org/news-insights/publication/mitres-privacy-engineering-tools-and-their-use-privacy-assessment.

[73] Government of New Zealand, "Privacy Maturity Assessment Framework (PMAF) and Self-Assessments," last updated 2021, accessed September 2025, www.digital.govt.nz/standards-and-

guidance/privacy-security-and-risk/privacy/privacy-maturity-assessment-framework-pmaf-and-self-assessments.

[74] Veronica Cretu, "Future-Ready Public Institutions: Rethinking Data Governance Through Maturity Assessment," *Economy and Sociology*, no. 1 (August 2025), https://doi.org/10.36004/nier.es.2025.1-04.

[75] D. L. Coates and A. Martin, "An Instrument to Evaluate the Maturity of Bias Governance Capability in Artificial Intelligence Projects," *IBM Journal of Research and Development* 63, no. 4/5 (2019): 1–15, https://doi.org/10.1147/JRD.2019.2915062.

[76] Angelantonio Castelli, Georgios Christos Chouliaras, and Dmitri Goldenberg, "Maturity Framework for Enhancing Machine Learning Quality," in *KDD '25: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2025), 4296–4307, https://doi.org/10.1145/3711896.3737246.

[77] Evidently AI, "Model Monitoring for ML in Production: A Aomprehensive Guide," Last updated: January 25, 2025, https://www.evidentlyai.com/ml-in-production/model-monitoring.

[78] "TMMi Model," TMMi Foundation, accessed August 2025, www.tmmi.org/tmmi-model/.

[79] Henri Sohier, Jean-Philippe Faure, Sebastian Hallensleben, Philippe Streiff, Stephen Creff, and Emilie Hien, "The Engineering of AI Evaluation and Scoring: Overview and Insights," in *2025 IEEE International Systems Conference (SysCon)* (IEEE, 2025), 1–8, https://doi.org/10.1109/SysCon64521.2025.11014820.

[80] Ahmad Al Mohamad Saleh and Saeed Alzahrani, "Development of a Maturity Model for Software Quality Assurance Practices," *Systems* 11, no. 9 (2023): 464, https://doi.org/10.3390/systems11090464.

[81] Joris Krijger, Tamara Thuis, Maarten de Ruiter, E. Ligthart, and I. Broekman, "The AI Ethics Maturity Model: A Holistic Approach to Advancing Ethical Data Science in Organizations," *AI Ethics* 3 (2023): 355–367, https://doi.org/10.1007/s43681-022-00228-7.

[82] Salesforce, "Salesforce Debuts AI Ethics Model: How Ethical Practices Further Responsible Artificial Intelligence," news release, September 2, 2021, www.salesforce.com/news/stories/salesforce-debuts-ai-ethics-model-how-ethical-practices-further-responsible-artificial-intelligence/.

[83] "Open Ethics Maturity Model," v1.0.0, Open Ethics, accessed August 2025, https://openethics.ai/oemm/.

[84] Vanja Skoric, Giovanni Sileno and Sennay Ghebreab, "Roles of Standardised Criteria in Assessing Societal Impact of AI," in *2024 IEEE Conference on Artificial Intelligence (CAI)* (IEEE, 2024), 1240–1245, https://doi.org/10.1109/CAI59869.2024.00220.

[85] Michael Mylrea and Nikki Robinson, "Artificial Intelligence (AI) Trust Framework and Maturity Model: Applying an Entropy Lens to Improve Security, Privacy, and Ethical AI," *Entropy* 25, no. 10 (2023): 1429, https://doi.org/10.3390/e25101429.

[86] Julián Muñoz-Ordóñez, Carlos Cobos, Juan C. Vidal-Rojas, and Francisco Herrera, "A Maturity Model for Practical Explainability in Artificial Intelligence-Based Applications: Integrating Analysis and Evaluation (MM4XAI-AE) Models," *International Journal of Intelligent Systems* 2025, no. 1 (2025): 1–18, https://doi.org/10.1155/int/4934696.

[87] "Organizational Context," Baldridge Performance Excellence Program, NIST, last modified June 5, 2023, www.nist.gov/baldrige/self-assessing/improvement-tools/foundations-successful-business/organizational-context.

[88] Dennis P. Watson, Erin L. Adams, Sarah Shue et al. "Defining the External Implementation Context: An Integrative Systematic Literature Review," *BMC Health Services Research* 18 (2018), https://doi.org/10.1186/s12913-018-3046-5.

[89] "The External Forces Influencing Business AI Governance," Edmond and Lily Safra Center for Ethics, Harvard University, March 1, 2024, www.ethics.harvard.edu/blog/post-7-external-forces-influencing-business-ai-governance%C2%A0.

[90] Jamie Johnson, "What Is a SWOT Analysis and How Do You Perform One?," U.S. Chamber of Commerce, June 4, 2025, www.uschamber.com/co/start/strategy/swot-analysis-guide.

[91] "PESTLE Analysis," Chartered Institute for Personnel and Development, March 21, 2025, www.cipd.org/uk/knowledge/factsheets/pestle-analysis-factsheet/.

[92] "Global AI Law and Policy Tracker," International Association of Privacy Professionals, accessed September 2025, https://iapp.org/resources/article/global-ai-legislation-tracker/.

[93] "AI Watch: Global Regulatory Tracker," White & Case, accessed September 2025, www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker.

[94] "Data Protection and Privacy Legislation Worldwide," United Nations Trade and Development, accessed September 2025, https://unctad.org/page/data-protection-and-privacy-legislation-worldwide.

[95] Pierre Sarlième, James Drummond, Becky King, Natalie Cohen, and Anna Pietikäinen, *A Mapping Tool for Digital Regulatory Frameworks: Including a Pilot on Efforts to Regulate AI* (OECD, 2025), https://dx.doi.org/10.1787/1cdad902-en.

[96] Okhaide Akhigbe, Daniel Amyot, and Gregory Richards, "A Systematic Literature Mapping of Goal and Non-goal Modelling Methods for Legal and Regulatory Compliance," *Requirements Engineering* 24 (2019): 459–481, https://doi.org/10.1007/s00766-018-0294-1.

[97] "Elicitation Techniques," Business Analysis, University of Notre Dame, accessed September 2025, https://sites.nd.edu/businessanalysis/?page_id=321.

[98] "Risk Appetite and Tolerance," Institute of Risk Management, accessed August 2025, www.theirm.org/what-we-say/thought-leadership/risk-appetite-and-tolerance/.

[99] Mary Carmichael, "Applying Risk Appetite and Risk Tolerance in the Age of AI," ISACA, August 19, 2024, www.isaca.org/resources/news-and-trends/newsletters/atisaca/2024/volume-16/applying-risk-appetite-and-risk-tolerance-in-the-age-of-ai.

[100] Bedanta Bora, Samarjeet Borah, and Wangchuk Chungyalpa, "Crafting Strategic Objectives: Examining the Role of Business Vision and Mission Statements," *Journal of Entrepreneurship Organization Management* 6 (2017), www.researchgate.net/publication/317799966_Crafting_Strategic_Objectives_Examining_the_Role_of_Business_Vision_and_Mission_Statements.

[101] Daniel Cochran, Fred David, and C. Kendrick Gibson, "A Framework for Developing an Effective Mission Statement," *Journal of Business Strategies* 25, no. 2 (2008): 27–39, https://jbs-ojs-shsu.tdl.org/jbs/article/view/133.

[102] Dalia Susnienė and Povilas Vanagas, "Development of Stakeholder Relationships by Integrating Their Needs into Organization's Goals and Objectives," *Engineering Economics* 48, no. 3 (2006): 83–87, www.ceeol.com/search/article-detail?id=29818.

[103] Robert S. Kaplan and David P. Norton, "The Balanced Scorecard—Measures that Drive Performance," *Harvard Business Review*, January–February 1992, https://hbr.org/1992/01/the-balanced-scorecard-measures-that-drive-performance-2.

[104] University of California, *SMART Goals: A How to Guide* (University of Michigan, accessed September 2025), https://hr.umich.edu/university-california-smart-goals-how-guide.

[105] "Aligning AI Initiatives with Business Objectives," RTS Labs, November 25, 2024, https://rtslabs.com/aligning-ai-initiatives-with-business-objectives.

[106] "Identifying and Scaling AI Use Cases," OpenAI, accessed September 2025, https://openai.com/business/guides-and-resources/identifying-and-scaling-ai-use-cases/.

[107] Tupokigwe Isagah and Soumaya I Ben Dhaou, "Problem Formulation and Use Case Identification of AI in Government: Results from the Literature Review," in *dg.o '23: Proceedings of the 24th Annual International Conference on Digital Government Research* (Association for Computing Machinery, 2023), 434–439, https://doi.org/10.1145/3598469.3598518.

[108] Matthias Brunnbauer, "Methods and Models for the Identification and Evaluation of AI Use Cases" (PhD diss., Johannes Gutenberg-Universität Mainz, 2024), https://openscience.ub.uni-mainz.de/server/api/core/bitstreams/ccd71ba5-5cc6-46ca-ba00-b0018b5a09b9/content.

[109] A. Leone de Castris, S. Laher, and F. Ostmann, *Business Applications of Artificial Intelligence - A Framework to Categorise AI Use Cases* (BridgeAI, 2025), https://doi.org/10.5281/zenodo.14727116.

[110] UK Department for Science, Innovation and Technology, Office for Artificial Intelligence, and Centre for Data Ethics and Innovation, "Assessing if Artificial Intelligence is the Right Solution," Gov.uk, June 10, 2019, www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution.

[111] Lynn Humpert, Moritz Wäschle, Sarah Horstmeyer, Harald Anacker, Roman Dumitrescu, and Albert Albers, "Stakeholder-Oriented Elaboration of Artificial Intelligence Use Cases using the Example of Special-Purpose Engineering," *Procedia CIRP* 119 (2023): 693–698, https://doi.org/10.1016/j.procir.2023.02.160.

[112] Holly J. Gregory and Sidley Austin, "AI and the Role of the Board of Directors," Harvard Law School Forum on Corporate Governance, October 7, 2023, https://corpgov.law.harvard.edu/2023/10/07/ai-and-the-role-of-the-board-of-directors/.

[113] Lara Abrash, Arno Probst, Karen Edelman, and Clare Harding, "Governance of AI: A Critical Imperative for Today's Boards," Deloitte, October 7, 2024, www.deloitte.com/us/en/insights/topics/leadership/successful-ai-oversight-may-require-more-engagement-in-the-boardroom.html.

[114] Peter Weill, Stephanie L. Woerner, Jennifer Banner, and James Moore, "Digitally Savvy Boards: AI Update," MIT Center for Information Systems Research, March 20, 2025, https://cisr.mit.edu/publication/2025_0301_SavvyBoardsUpdate_WeillWoernerBannerMoore.

[115] National Association of Corporate Directors and Data and Trust Alliance, "AI and Board Governance," NACD, September 18, 2023, www.nacdonline.org/all-governance/governance-resources/governance-research/director-faqs-and-essentials/ai-and-board-governance/.

[116] "Cyber Governance for Boards," UK National Cyber Security Centre, accessed September 2025, www.ncsc.gov.uk/cyber-governance-for-boards/overview.

[117] John Winsor, Jen Stave, and Ryan Kurt, "Your AI Strategy Needs More Than a Single Leader," *Harvard Business Review*, August 4, 2025, https://hbr.org/2025/08/your-ai-strategy-needs-more-than-a-single-leader.

[118] Marc Schmitt, "Strategic Integration of Artificial Intelligence in the C-Suite: The Role of the Chief AI Officer," arXiv preprint arXiv:2407.10247 (2024), https://doi.org/10.48550/arXiv.2407.10247.

[119] Mathias Schäfer, Johannes Schneider, Katharina Drechsler, Jan vom Brocke, "AI Governance: Are Chief AI Officers and AI Risk Officers Needed?," in *2022 ECIS Proceedings* (ECIS, May 2022), https://aisel.aisnet.org/ecis2022_rp/163.

[120] Michael Wade, Anja Lagodny, Ann-Christin Andersen, Corinne Avelines, and Achim Plueckebaum, "Do You Really Need a Chief AI Officer?," *MIT Sloan Management Review*, August 7, 2024, https://sloanreview.mit.edu/article/do-you-really-need-a-chief-ai-officer/.

[121] Yousuf Alblooshi, Amin Hosseinian-Far, and Dilshad Sarwar, "Identification of Critical Business Processes: A Proposed Novel Approach," in *Cybersecurity, Privacy and Freedom Protection in the Connected World* (Springer, 2021), https://doi.org/10.1007/978-3-030-68534-8_25.

[122] James J. Cebula, Mary E. Popeck, and Lisa R. Young, "A Taxonomy of Operational Cyber Security Risks Version 2," Software Engineering Institute, Carnegie Mellon University, May 2014, www.sei.cmu.edu/library/a-taxonomy-of-operational-cyber-security-risks-version-2/.

[123] European Union Agency for Cybersecurity, *Reference Incident Classification Taxonomy: Task Force Status and Way Forward* (ENISA, January 2018), www.enisa.europa.eu/publications/reference-incident-classification-taxonomy.

[124] "A Taxonomy of Privacy," Open Rights Group Wiki, accessed October 2025, https://wiki.openrightsgroup.org/wiki/A_Taxonomy_of_Privacy.

[125] Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy et al., "The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence," arXiv preprint arXiv:2408.12622 (2025), https://doi.org/10.48550/arXiv.2408.12622.

[126] Hao-Ping (Hank) Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das, "Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks," in *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2024), https://doi.org/10.1145/3613904.3642116.

[127] Renee Shelby, Shalaleh Rismani, Kathryn Henne et al., "Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction," in *AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery, 2023), 723–741, https://doi.org/10.1145/3600211.3604673.

[128] ORX and Olvier Wyman, *The ORX Reference Taxonomy for Operational and Non-financial Risk* (ORX, 2025), https://orx.org/download/orx-reference-taxonomy.

[129] George Grispos, William Bradley Glisson, and Tim Storer, "Security Incident Response Criteria: A Practitioner's Perspective," arXiv preprint, arXiv:1508.02526 (2025), https://doi.org/10.48550/arXiv.1508.02526.

[130] "Defining an Incident" in *The Tactical Playbook for Modern Incident Management* (Incident.io, 2025), https://incident.io/guide/foundations/defining-an-incident.

[131] ServiceDesk Plus, "What is an ITIL® priority matrix", January 24, 2024, https://www.manageengine.com/products/service-desk/itsm/itil-priority-matrix.html.

[132] "Resources and Constraints," Canada Energy Regulator, last modified November 3, 2023, www.cer-rec.gc.ca/en/safety-environment/safety-culture/safety-culture-learning-portal/human-organizational-factors/resources-constraints/.

[133] Ben Buchanan, *The AI Triad and What It Means for National Security Strategy* (CSET, August 2020), https://doi.org/10.51593/20200021.

[134] Micah Musser, Rebecca Gelles, Ronnie Kinoshita, Catherine Aiken, and Andrew Lohn, *The Main Resource Is the Human: A Survey of AI Researchers on the Importance of Compute* (CSET, April 2023), https://doi.org/10.51593/20210071.

[135] Kiran A. Ahuja, "The AI in Government Act of 2020—Artificial Intelligence Competencies," memorandum, U.S. Office of Personnel Management (OPM), July 6, 2023, www.opm.gov/chcoc/published-memos/.

[136] Patricia Caratozzolo, Jose Daniel Azofeifa, Luis Alberto Mejía-Manzano, Valentina Rueda-Castro, Julieta Noguez, and Alejandra J. Magana, "A Matrix Taxonomy of Knowledge, Skills, and Abilities (KSA) Shaping 2030 Labor Market," in *2023 IEEE Frontiers in Education Conference* (IEEE, 2023), 1–8, https://doi.org/10.1109/FIE58773.2023.10342955.

[137] Southern Regional Education Board Commission on Artificial Intelligence in Education, *Skills for an AI-Ready Workforce* (SREB, July 2025), www.sreb.org/publication/skills-ai-ready-workforce-0.

[138] Ania W. Masinter, "Assessing Your Talent Needs in the Age of AI," *Harvard Business Review*, May 28, 2025, https://hbr.org/2025/05/assessing-your-talent-needs-in-the-age-of-ai.

[139] IT Modernization Center of Excellence, "Develop the AI Workforce," in *AI Guide for Government* (U.S. General Services Administration, 2025), ch. 4, https://coe.gsa.gov/coe/ai-guide-for-government/developing-ai-workforce/index.html.

[140] UTS, "Harnessing Companies' People, Skills and Culture for Effective AI Use," news release, December 2, 2024, www.uts.edu.au/news/2024/12/harnessing-companies-people-skills-and-culture-effective-ai-use.

[141] Diana Gehlhaus and Santiago Mutis, *The U.S. AI Workforce: Understanding the Supply of AI Talent* (CSET, January 2021), https://doi.org/10.51593/20200068.

[142] Sonali Subbu Rathinam, "The U.S. AI Workforce: Analyzing Current Supply and Growth," CSET, January 30, 2024, https://cset.georgetown.edu/publication/the-u-s-ai-workforce-analyzing-current-supply-and-growth/.

[143] Autumn Toney and Melissa Flagg, *U.S. Demand for AI-Related Talent* (CSET, August 2020), https://doi.org/10.51593/20200027.

[144] Nash Squared and Harvey Nash, *Digital Leadership Report 2025* (Harvey Nash, 2025), www.harveynash.co.uk/research-whitepapers/digital-leadership-report-2025.

[145] William Newhouse, Murugiah Souppaya, John Kent, Ken Sandlin, and Karen Scarfone, *Data Classification Concepts and Considerations for Improving Data Protection* (NIST, November 2023), https://doi.org/10.6028/NIST.IR.8496.ipd.

[146] Catherine Aiken, *Classifying AI Systems* (CSET, November 2021), https://doi.org/10.51593/20200025.

[147] *OECD Framework for the Classification of AI Systems* (OECD, February 2022), https://oecd.ai/en/classification.

[148] Anas Baig and Ozair Malik, "What Is Data Classification Policy? Example and Templates Included," Securiti, December 3, 2024, https://securiti.ai/data-classification-policy/.

[149] "Guidance on Legal Bases for Processing Personal Data," Irish Data Protection Commission, accessed September 2025, www.dataprotection.ie/en/dpc-guidance/guidance-legal-bases-processing-personal-data.

[150] "A Guide to Lawful Basis," UK Information Commissioner's Office, accessed September 2025, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/a-guide-to-lawful-basis/.

[151] "Ensuring the Lawfulness of the Data Processing—Defining a Legal Basis," CNIL, June 7, 2024, www.cnil.fr/en/ensuring-lawfulness-data-processing-legal-basis.

[152] Johanna Rothman, "Release Criteria: Good to Go" (Villanova University, 2007), www.researchgate.net/profile/Johanna-Rothman/publication/237830411_Release_Criteria_Good_to_Go_Part_1_Using_Release_Criteria_to_Check_Your_Software_for_Doneness/links/5422e8620cf238c6ea6e2c69/Release-Criteria-Good-to-Go-Part-1-Using-Release-Criteria-to-Check-Your-Software-for-Doneness.pdf.

[153] Solon Barocas, Asia J. Biega, Benjamin Fish, Jędrzej Niklas, Luke Stark, "When Not to Design, Build, or Deploy," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2020), https://doi.org/10.1145/3351095.3375691.

[154] "Risk Taxonomy and Thresholds for Frontier AI Frameworks," Frontier Model Forum, June 18, 2025, www.frontiermodelforum.org/technical-reports/risk-taxonomy-and-thresholds/.

[155] Jonas Schuett, Eunseo Choi, Kasumi Sugimoto, Bosco Hung, Robert Trager, and Karine Perset, *Survey on Thresholds for Advanced AI Systems* (AI Governance Initiative, August 2025), https://aigi.ox.ac.uk/publications/survey-on-thresholds-for-advanced-ai-systems/.

[156] Deepika Raman, Nada Madkour, Evan R. Murphy, Krystal Jackson, and Jessica Newman, *Intolerable Risk Threshold Recommendations for Artificial Intelligence* (Center for Long-term Cybersecurity, February 2025), https://cltc.berkeley.edu/publication/intolerable-ai-risk-thresholds/.

[157] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao, "The Values Encoded in Machine Learning Research," in *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2022), 173–184, https://doi.org/10.1145/3531146.3533083.

[158] Jenna L. Butler, Thomas Zimmermann, and Christian Bird, "Objectives and Key Results in Software Teams: Challenges, Opportunities and Impact on Development," in *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice* (Association for Computing Machinery, 2024), 358–368. https://doi.org/10.1145/3639477.3639747.

[159] Didar Zowghi and Chad Coulin, "Requirements Elicitation: A Survey of Techniques, Approaches, and Tools," in *Engineering and Managing Software Requirements* (Springer, 2005), https://doi.org/10.1007/3-540-28244-0_2.

[160] Carla Pacheco and Ivan Garcia, "A Systematic Literature Review of Stakeholder Identification Methods in Requirements Elicitation," *Journal of Systems and Software* 85, no. 9 (2012): 2171–2181, https://doi.org/10.1016/j.jss.2012.04.075.

[161] Fahim Muhammad Khan, Javed Ali Khan, Muhammad Assam, Ahmed S. Almasoud, Abdelzahir Abdelmaboud, Manar Ahmed Mohammed Hamza, "A Comparative Systematic Analysis of Stakeholder's Identification Methods in Requirements Elicitation," *IEEE Access* 10 (2022): 30982–31011, https://doi.org/10.1109/ACCESS.2022.3152073.

[162] Diana Robinson, Christian Cabrera, Andrew D. Gordon, Neil D. Lawrence, and Lars Mennen, "Requirements Are All You Need: The Final Frontier for End-User Software Engineering," *ACM Transactions on Software Engineering and Methodology* 34, no. 5 (June 2025): 1–22, https://doi.org/10.1145/3708524.

[163] Khlood Ahmad, Mohamed Abdelrazek, Chetan Arora, Muneera Bano, and John Grundy, "Requirements Engineering for Artificial Intelligence Systems: A Systematic Mapping Study," *Information and Software Technology* 158 (2023), https://doi.org/10.1016/j.infsof.2023.107176.

[164] Tor Sporsem, Rasmus Ulfsnes, Morten Hatling, and Inga Strümke, "Clash of Requirements: Users First vs. Model First," in *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering* (Association for Computing Machinery, 2025), 1515–1519, https://doi.org/10.1145/3696630.3731668.

[165] Nick Feng, Lina Marsso, Sinem Getir Yaman, Beverley Townsend, Ana Cavalcanti, Radu Calinescu, "Towards a Formal Framework for Normative Requirements Elicitation," in *Proceedings of the 38th IEEE/ACM International Conference on Automated Software Engineering* (IEEE, 2024): 1776–1780, https://doi.org/10.1109/ASE56229.2023.00152.

[166] "Catalogue of Tools and Metrics for Trustworthy AI," OECD.AI Policy Observatory, accessed September 2025, https://oecd.ai/en/catalogue/metrics.

[167] "Measuring Gen AI Success: A Deep Dive into the KPIs You Need," Google, November 25, 2024, https://cloud.google.com/transform/gen-ai-kpis-measuring-ai-success-deep-dive.

[168] "Observability in Generative AI," Microsoft, August 2, 2025), https://learn.microsoft.com/en-us/azure/ai-foundry/concepts/observability.

[169] Abigail Z. Jacobs and Hanna Wallach, "Measurement and Fairness," *in FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2021), 375–385, https://doi.org/10.1145/3442188.3445901.

[170] Md Abdul Kadir, Amir Mosavi, and Daniel Sonntag, "Evaluation Metrics for XAI: A Review, Taxonomy, and Practical Applications," *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)* (IEEE, 2023), 111–124, https://doi.org/10.1109/INES59282.2023.10297629.

[171] Isabel Wagner and David Eckhoff, "Technical Privacy Metrics: A Systematic Survey," *ACM Computing Surveys* 51, no. 3 (May 2019): 1–38, https://doi.org/10.1145/3168389.

[172] Tamar Eilam, Pedro Bello-Maldonado, Bishwaranjan Bhattacharjee, Carlos Costa, Eun Kyung Lee, and Asser Tantawi, "Towards a Methodology and Framework for AI Sustainability Metrics," *HotCarbon '23: Proceedings of the 2nd Workshop on Sustainable Computer Systems* (Association for Computing Machinery, 2023), https://doi.org/10.1145/3604930.3605715.

[173] Steven A. Melnyk, Roger J. Calantone, Joan Luft et al., "An Empirical Investigation of the Metrics Alignment Process," *International Journal of Productivity and Performance Management* 54, no. 5-6 (2005): 312–324, https://doi.org/10.1108/17410400510604494.

[174] "AI Capabilities," Epoch AI, last modified March 11, 2026, https://epoch.ai/benchmarks.

[175] Kathrin Blagec, Georg Dorffner, Milad Moradi, and Matthias Samwald, "A Critical Analysis of Metrics Used for Measuring Progress in Artificial Intelligence," arXiv preprint arXiv:2008.02577 (2021), https://doi.org/10.48550/arXiv.2008.02577.

[176] Richard Ren, Steven Basart, Adam Khoja et al., "Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?," *Advances in Neural Information Processing Systems* 37 (2024): 68559–68594, https://proceedings.neurips.cc/paper_files/paper/2024/hash/7ebcdd0de471c027e67a11959c666d74-Abstract-Datasets_and_Benchmarks_Track.html.

[177] Salih Caner and Feyza Bhatti, "A Conceptual Framework on Defining Businesses Strategy for Artificial Intelligence," *Contemporary Management Research* 16, no. 3 (2020): 175–206, https://doi.org/10.7903/cmr.19970.

[178] Angelina Wang, Teresa Datta, and John P. Dickerson, "Strategies for Increasing Corporate Responsible AI Prioritization," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (October 16, 2024): 1514–1526, https://doi.org/10.1609/aies.v7i1.31743.

[179] Marios Constantinides, Edyta Bogucka, Daniele Quercia, Susanna Kallio, and Mohammad Tahaei, "RAI Guidelines: Method for Generating Responsible AI Guidelines Grounded in Regulations and Usable by (Non-)Technical Roles," *Proceedings of ACM Human-Computer Interaction* 8 (November 2024): 1–28, https://doi.org/10.1145/3686927.

[180] Nick Malter, "Writing an Organizational AI Policy: First Step Towards Effective AI Governance," Apply AI Alliance, September 17, 2024, https://futurium.ec.europa.eu/en/european-ai-alliance/community-content/writing-organizational-ai-policy-first-step-towards-effective-ai-governance.

[181] "Build Your Foundational Organizational AI Policy," Responsible Artificial Intelligence Institute, accessed, August 2025, www.responsible.ai/ai-policy-template/.

[182] Mary Carmichael, "Key Considerations for Developing Organizational Generative AI Policies," ISACA, November 1, 2023, www.isaca.org/resources/news-and-trends/newsletters/atisaca/2023/volume-44/key-considerations-for-developing-organizational-generative-ai-policies.

[183] Kevin Klyman, "Acceptable Use Policies for Foundation Models," *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (2025): 752–767, https://doi.org/10.1609/aies.v7i1.31677.

[184] Amber Ezzell, *Generative AI for Organizational Use: Internal Policy Checklist* (Future of Privacy Forum, July 2023), https://fpf.org/blog/fpf-releases-generative-ai-internal-policy-checklist-to-guide-development-of-policies-to-promote-responsible-employee-use-of-generative-ai-tools/.

[185] Catherine A. Maritan and Gwendolyn K. Lee, "Resource Allocation and Strategy," *Journal of Management* 43, no. 8 (2017): 2411–2420, https://doi.org/10.1177/0149206317729738.

[186] Noman Bashir, Priya Donti, James Cuff et al., "The Climate and Sustainability Implications of Generative AI," MIT, March 27, 2024, https://mit-genai.pubpub.org/pub/8ulgrckc/release/2.

[187] IT Modernization Center of Excellence, "Understanding and Managing the AI Lifecycle," in *AI Guide for Government* (U.S. General Services Administration, 2025), ch. 7, https://coe.gsa.gov/coe/ai-guide-for-government/understanding-managing-ai-lifecycle/.

[188] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, and Ece Kamar, "Software Engineering for Machine Learning: A Case Study," *in 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice* (IEEE, 2019), 291–300, https://doi.org/10.1109/ICSE-SEIP.2019.00042.

[189] *ISO/IEC/IEEE 12207:2017 Systems and Software Engineering—Software Life Cycle Processes*, (ISO, 2017), www.iso.org/standard/63712.html.

[190] Daniel Lübke, "Design Patterns for Approval Processes," in *Proceedings of the 28th European Conference on Pattern Languages of Programs* (Association for Computing Machinery, 2023), https://doi.org/10.1145/3628034.3628035.

[191] Silverio Martínez-Fernández, Justus Bogner, Xavier Franch et al., "Software Engineering for AI-Based Systems: A Survey," *ACM Transactions on Software Engineering Methodology* 31, no. 2 (April 2022): 1–59, https://doi.org/10.1145/3487043.

[192] Blaine Kuehnert, Rachel Kim, Jodi Forlizzi, and Hoda Heidari, "The 'Who,' 'What,' and 'How' of Responsible AI Governance: A Systematic Review and Meta-Analysis of (Actor, Stage)-Specific Tools," in *FAccT '25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, June 2025), 2991–3005, https://doi.org/10.1145/3715275.3732191.

[193] "Cyber Security Incident Response Planning: Practitioner Guidance," Australian Signals Directorate, last modified December 12, 2024, www.cyber.gov.au/business-government/detecting-responding-to-threats/cyber-security-incident-response/cyber-security-incident-response-planning-practitioner-guidance.

[194] "Plan: Your Cyber Incident Response Processes," Incident Management, UK National Cyber Security Centre, September 19, 2019, www.ncsc.gov.uk/collection/incident-management/cyber-incident-response-processes.

[195] *Cybersecurity Incident and Vulnerability Response Playbooks: Operational Procedures for Planning and Conducting Cybersecurity Incident and Vulnerability Response Activities in FCEB Information System* (CISA, November 2021), www.cisa.gov/resources-tools/resources/federal-government-cybersecurity-incident-and-vulnerability-response-playbooks.

[196] *GenAI Incident Response Guide* (OWASP Foundation, July 28, 2025), https://genai.owasp.org/resource/genai-incident-response-guide-1-0/.

[197] "Business Impact Analysis," Ready.gov, U.S. Department of Homeland Security, last modified December 26, 2023, www.ready.gov/business/planning/impact-analysis.

[198] *Infrastructure Resilience Planning Framework* (CISA, March 2025), www.cisa.gov/resources-tools/resources/infrastructure-resilience-planning-framework-irpf.

[199] Derrick Musundi Kesa, "Ensuring Resilience: Integrating IT Disaster Recovery Planning and Business Continuity for Sustainable Information Technology Operations," *World Journal of Advanced Research and Reviews* 18, no. 3 (2023): 970–999, https://doi.org/10.30574/wjarr.2023.18.3.1166.

[200] "Systems Engineering: Risk Mitigation," MITRE, accessed September 2025, www.mitre.org/our-impact/mitre-labs/systems-engineering-innovation-center/risk-mitigation.

[201] Kiran A. Ahuja, "Skills-Based Hiring Guidance and Competency Model for Artificial Intelligence Work," memorandum, OPM, April 29, 2024, www.opm.gov/chcoc/published-memos/.

[202] OPM, *Workforce Planning Guide* (November 2022), www.opm.gov/chcoc/published-memos/.

[203] Jose Daniel Azofeifa, Luis Jose Gonzalez-Gomez, Valentina Rueda-Castro, Sonia M. Gómez-Puente, Julieta Noguez, Patricia Caratozzolo, "Top Occupations Based on a Strategic Taxonomy Framework of Future Skills for Workforce Development," in *2025 IEEE Global Engineering Education Conference* (IEEE, 2025): 1–8, https://doi.org/10.1109/EDUCON62633.2025.11016445.

[204] Matthias Oschinski, Ali Crawford, and Maggie Wu, *AI and the Future of Workforce Training* (CSET, December 2024), https://cset.georgetown.edu/publication/ai-and-the-future-of-workforce-training/.

[205] Michaela Poláková, Juliet Horváthová Suleimanová, Peter Madzík, Lukáš Copuš, Ivana Molnárová, and Jana Polednová, "Soft Skills and Their Importance in the Labour Market Under the

Conditions of Industry 5.0," *Heliyon* 9, no. 8 (July 27, 2023), https://doi.org/10.1016/j.heliyon.2023.e18670.

[206] Martha Gimbel, Molly Kinder, Joshua Kendall, and Maddie Lee, "Evaluating the Impact of AI on the Labor Market: Current State of Affairs," Yale Budget Lab, October 1, 2025, https://budgetlab.yale.edu/research/evaluating-impact-ai-labor-market-current-state-affairs.

[207] Carlo Pizzinelli, Augustus J Panton, Marina Mendes Tavares, Mauro Cazzaniga, and Longji Li, *Labor Market Exposure to AI: Cross-Country Differences and Distributional Implications* (International Monetary Fund, October 4, 2023), www.imf.org/en/Publications/WP/Issues/2023/10/04/Labor-Market-Exposure-to-AI-Cross-country-Differences-and-Distributional-Implications-539656.

[208] Rakesh Kochhar, "Which U.S. Workers Are More Exposed to AI on Their Jobs," Pew Research Center, July 26, 2023, www.pewresearch.org/social-trends/2023/07/26/which-u-s-workers-are-more-exposed-to-ai-on-their-jobs/.

[209] Edward Felten, Manav Raj, and Robert Seamans, "Occupational, Industry, and Geographic Exposure to Artificial Intelligence: A Novel Dataset and Its Potential Uses," *Strategic Management Journal* 42 (2021): 2195–2217, https://doi.org/10.1002/smj.3286.

[210] Tracy Mayor, "Ethics and Automation: What to Do When Workers Are Displaced," *Ideas Made to Matter* (blog), MIT Sloan School of Management, July 8, 2019, https://mitsloan.mit.edu/ideas-made-to-matter/ethics-and-automation-what-to-do-when-workers-are-displaced.

[211] "From Adoption to Empowerment: Shaping the AI-Driven Workforce of Tomorrow," Society for Human Resource Management, July 8, 2025, www.shrm.org/topics-tools/research/from-adoption-to-empowerment--shaping-the-ai-driven-workforce-of-tomorrow.

[212] Jorge Tamayo, Leila Doumi, Sagar Goel, Orsolya Kovács-Ondrejkovic, and Raffaella Sadun, "Reskilling in the Age of AI: Five New Paradigms for Leaders—and Employees," *Harvard Business Review*, September–October 2023, https://hbr.org/2023/09/reskilling-in-the-age-of-ai.

[213] Julian Jacobs, "AI Labor Displacement and the Limits of Worker Retraining," Brookings, May 16, 2025, www.brookings.edu/articles/ai-labor-displacement-and-the-limits-of-worker-retraining/.

[214] Azad M. Madni and Michael Sievers, "Systems Integration: Key Perspectives, Experiences, and Challenges," *Systems Engineering* 17 (2013): 37–51, https://doi.org/10.1002/sys.21249.

[215] Ramaswamy Chandramouli and Doron Pinhas, *Security Guidelines for Storage Infrastructure* (NIST, October 2020), https://doi.org/10.6028/NIST.SP.800-209.

[216] Matt Pacheco, "Essential Considerations for Data Storage Capacity Planning," TierPoint, last modified December 11, 2025, www.tierpoint.com/blog/data-storage-capacity-planning/.

[217] Adebimpe Bolatito Ige, Naomi Chukwurah, Courage Idemudia, and Victor Ibukun Adebayo, "Managing Data Lifecycle Effectively: Best Practices for Data Retention and Archival Processes,"

International Journal of Engineering Research and Development 20, no. 7 (July 2024): 453–461, www.ijerd.com/paper/vol20-issue7/.

[218] "What Is a Data Retention Policy? Best Practices and Template," Drata, August 12, 2025, https://drata.com/blog/data-retention-policy.

[219] Yupeng Chang, Xu Wang, Jindong Wang et al., "A Survey on Evaluation of Large Language Models," *ACM Transactions on Intelligent Systems and Technology* 15, no. 3 (June 2024): 1–45, https://doi.org/10.1145/3641289.

[220] Chief Digital and Artificial Intelligence Office, *Test and Evaluation of Artificial Intelligence Models: What to Consider in a Test & Evaluation Strategy* (U.S. Department of War, April 2024), www.ai.mil/Latest/Blog/Article-Display/Article/3940283/cdao-test-and-evaluation-strategy-frameworks/.

[221] John Burden, Marko Tešić, Lorenzo Pacchiardi, and José Hernández-Orallo, "Paradigms of AI Evaluation: Mapping Goals, Methodologies and Culture," arXiv preprint arXiv:2502.15620 (2025), https://doi.org/10.48550/arXiv.2502.15620.

[222] Jonathan Spring and Divjot Singh Bawa, "AI Red Teaming: Applying Software TEVV for AI Evaluations," CISA, November 26, 2024, www.cisa.gov/news-events/news/ai-red-teaming-applying-software-tevv-ai-evaluations.

[223] Evelyn Yee, "AI Red-Teaming Design: Threat Models and Tools," CSET, October 24, 2025, https://cset.georgetown.edu/article/ai-red-teaming-design-threat-models-and-tools.

[224] Irene Solaiman, Rishi Bommasani, Dan Hendrycks et al., "Beyond Release: Access Considerations for Generative AI Systems," arXiv preprint arXiv:2502.16701 (2025), https://doi.org/10.48550/arXiv.2502.16701.

[225] Ashwin Acharya and Oscar Delaney, *Managing Risks from Internal AI Systems* (Institute for AI Policy and Strategy, July 2025), www.iaps.ai/research/managing-risks-from-internal-ai-systems.

[226] Irene Solaiman, "The Gradient of Generative AI Release: Methods and Considerations," in *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2023), 111–122, https://doi.org/10.1145/3593013.3593981.

[227] Elizabeth Seger, Noemi Dreksler, Richard Moulange et al., *Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives* (Centre for the Governance of AI, September 2023), www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models.

[228] Luis Aranda and Karine Perset, "AI Openness: Balancing Innovation, Transparency and Risk in Open-Weight Models," OECD.AI Policy Observatory, August 28, 2025, https://oecd.ai/en/wonk/balancing-innovation-transparency-and-risk-in-open-weight-models.

[229] Jon Bateman, Dan Baer, Stephanie A. Bell et al., "Beyond Open vs. Closed: Emerging Consensus and Key Questions for Foundation AI Model Governance," Carnegie Endowment for International Peace, July 23, 2024, https://carnegieendowment.org/research/2024/07/beyond-open-vs-closed-emerging-consensus-and-key-questions-for-foundation-ai-model-governance.

[230] Kristin Burnham, "Buy, Boost, or Build? Choose Your Path to Generative AI," *Ideas Made to Matter* (blog), MIT Sloan School of Management, September 17, 2025, https://mitsloan.mit.edu/ideas-made-to-matter/buy-boost-or-build-choose-your-path-to-generative-ai.

[231] Jeremy Licata, Rebecca McWhite, Laura Calloway et al., *Developing Security, Privacy, and Cybersecurity Supply Chain Risk Management Plans for Systems* (NIST, June 2025), https://doi.org/10.6028/NIST.SP.800-18r2.ipd.

[232] Anka Reuel, Benjamin Bucknall, Stephen Casper et al., "Open Problem in Technical AI Governance," arXiv preprint arXiv:2407.14981 (2004), https://doi.org/10.48550/arXiv.2407.14981.

[233] "Introduction," DevIQ, accessed October 2025, https://deviq.com/.

[234] Chip Huyen, *Designing Machine Learning Systems* (O'Reilly, May 2022), www.oreilly.com/library/view/designing-machine-learning/9781098107956/.

[235] Justin D. Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer, "Design Principles for Generative AI Applications" *in CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2024), https://doi.org/10.1145/3613904.3642466.

[236] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu et al., "Guidelines for Human-AI Interaction," in *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2019), https://doi.org/10.1145/3290605.3300233.

[237] "Intro to Deployment Strategies: Blue-Green, Canary, and More," Harness, January 15, 2021, www.harness.io/blog/blue-green-canary-deployment-strategies.

[238] Sydney Lesser, "What Is Ring Deployment? A Guide to Phase Software Rollouts," Ivanti, April 22, 2025, www.ivanti.com/blog/ring-deployment.

[239] "Risk Culture," Institute of Risk Management, accessed August 2025, www.theirm.org/what-we-say/thought-leadership/risk-culture/.

[240] Tim Fountaine, Brian McCarthy, and Tamim Saleh, "Building the AI-powered Organization," *Harvard Business Review*, July–August 2019, https://hbr.org/2019/07/building-the-ai-powered-organization.

[241] *Culture of AI Benchmark Report: State of AI Adoption and Culture Readiness in Europe* (Gallup, 2024), www.gallup.com/workplace/652784/culture-of-ai-and-adoption-report.aspx.

[242] David Manheim, "Building a Culture of Safety for AI: Perspectives and Challenges," SSRN, June 26, 2023, http://dx.doi.org/10.2139/ssrn.4491421.

[243] Emilia N. Mwim and Jabu Mtsweni, "Systematic Review of Factors that Influence the Cybersecurity Culture," *Human Aspects of Information Security and Assurance* 658 (2022), https://doi.org/10.1007/978-3-031-12172-2_12.

[244] Eko Yon Handri, Dana Indra Sensuse, and Sofian Lusa, "Examining Cybersecurity Culture: Trends and Success Factors," *Journal of Internet Services and Information Security* 14, no. 3 (2024): 330–352, https://doi.org/10.58346/JISIS.2024.I3.020.

[245] Moneer Alshaikh, "Developing Cybersecurity Culture to Influence Employee Behavior: A Practice Perspective," *Computers & Security* 98 (2020), https://doi.org/10.1016/j.cose.2020.102003.

[246] Leonardo Horn Iwaya, Gabriel Horn Iwaya, Simone Fischer-Hübner, and Andrea Valéria Steil, "Organisational Privacy Culture and Climate: A Scoping Review," *IEEE Access* 10 (2022): 73907–73930, https://doi.org/10.1109/ACCESS.2022.3190373.

[247] Muhammad Asif Qureshi, "Building a Privacy Culture," *ISACA Journal* (2 September 2, 2020), www.isaca.org/resources/isaca-journal/issues/2020/volume-5/building-a-privacy-culture.

[248] Mohammad Tahaei, Alisa Frik, and Kami Vaniea, "Privacy Champions in Software Teams: Understanding Their Motivations, Strategies, and Challenges," in *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2021), https://doi.org/10.1145/3411764.3445768.

[249] "The IBE Business Ethics Framework," Institute of Business Ethics, accessed September 2025, www.ibe.org.uk/knowledge-hub/ibe-business-ethics-framework.html.

[250] "Resources and Trainings for Ethics Officials," U.S. Office of Government Ethics, accessed September 2025, https://www.oge.gov/web/oge.nsf/section_landing_resources.

[251] Jonas Schuett, Ann-Katrin Reuel, and Alexis Carlier, "How to Design an AI Ethics Board," *AI and Ethics* 5 (2025): 863–881, https://doi.org/10.1007/s43681-023-00409-y.

[252] Emanuel Moss and Jacob Metcalf, *Ethics Owners: A New Model of Organizational Responsibility in Data-Driven Technology Companies* (Data & Society Research Institute, 2020), https://datasociety.net/library/ethics-owners/.

[253] *Recommendation on the Ethics of Artificial Intelligence* (UNESCO, 2022), https://unesdoc.unesco.org/ark:/48223/pf0000381137.

[254] Isaac H. Smith and Maryam Kouchaki, "Building an Ethical Company," *Harvard Business Review*, November–December 2021, https://hbr.org/2021/11/building-an-ethical-company.

[255] *OECD Due Diligence Guidance for Responsible Business Conduct* (OECD Publishing, 2018), https://doi.org/10.1787/15f5f4b3-en.

[256] Jacqueline de Gramont, *The Business Case for Speaking Up: How Internal Reporting Mechanisms Strengthen Private-Sector Organizations* (Transparency International, July 2017), www.transparency.org/en/publications/business-case-for-speaking-up.

[257] Amelia Lee Dogan, Hongjin Lin, and Lindah Kotut, "Down to Earth: Design Considerations for AI for Sustainability from the Environmental and Climate Movement," *DIS '25: Proceedings of the 2025 ACM Designing Interactive Systems Conference* (Association for Computing Machinery, 2025), 1549–1562, https://doi.org/10.1145/3715336.3735734.

[258] *Measuring the Environmental Impacts of Artificial Intelligence Compute and Applications: The AI Footprint* (OECD Publishing, November 2022), https://doi.org/10.1787/7babf571-en.

[259] Rohan Sharma, "AI Operating Model," in *AI and the Boardroom* (Apress, 2024), https://doi.org/10.1007/979-8-8688-0796-1_7.

[260] Tinglong Dai and Jayashankar M. Swaminathan, "AI and Operations: A Foundational Framework of Emerging Research and Practice," *Production and Operations Management*, https://dx.doi.org/10.2139/ssrn.5418934.

[261] *Guidance for Preparing Standard Operating Procedures (SOPs)* (U.S. Environmental Protection Agency, April 2007), www.epa.gov/sites/default/files/2015-06/documents/g6-final.pdf.

[262] *Guidance on the Core Principles of a Credible and Effective Compliance Program* (Competition Bureau, January 2024), https://competition-bureau.canada.ca/en/how-we-foster-competition/compliance-and-enforcement/core-principles-credible-and-effective-compliance-program.

[263] Rao Faizan Ali, P. D. D. Dominic, Syed Emad Azhar Ali, Mobashar Rehman, and Abid Sohail, "Information Security Behavior and Information Security Policy Compliance: A Systematic Literature Review for Identifying the Transformation Process from Noncompliance to Compliance," *Applied Sciences* 11, no. 8 (2021): 3383, https://doi.org/10.3390/app11083383.

[264] Shaobo Wei, Yuanyuan Zhang, and John Qi Dong, "Toward Artificial Intelligence Compliance: Impacts and Mechanisms of Performance Feedback," *Information Systems Research* (2025): 1–26, https://doi.org/10.1287/isre.2023.0580.

[265] David Jancsics, Salvador Espinosa, and Jonathan Carlos, "Organizational Noncompliance: An Interdisciplinary Review of Social and Organizational Factors," *Management Review Quarterly* 73 (2023): 1273–1301, https://doi.org/10.1007/s11301-022-00274-9.

[266] *How to Hold Employees Accountable: A Leader's Guide to Accountability in the Workplace* (Niagara Institute, 2022), www.niagarainstitute.com/blog/guide/accountability.

[267] Deborah Theseira, "How to Effectively Empower Teams with an Accountability Framework," Ardoq, July 7, 2022, www.ardoq.com/blog/accountability-framework.

[268] Jeanne W. Ross, Cynthia M. Beath, and Martin Mocker, *Designed for Digital: How to Architect Your Business for Sustained Success* (MIT Press, 2021), https://mitpress.mit.edu/9780262542760/designed-for-digital/.

[269] Ron Carucci, "How to Actually Encourage Employee Accountability," *Harvard Business Review*, November 23, 2020, https://hbr.org/2020/11/how-to-actually-encourage-employee-accountability.

[270] Peg Thoms, Jennifer J. Dose, and Kimberly S. Scott, "Relationships Between Accountability, Job Satisfaction, and Trust," *Human Resource Development Quarterly* 13 (September 16, 2002): 307–323, https://doi.org/10.1002/hrdq.1033.

[271] Virginia R. Stewart, Deirdre G. Snyder, and Chia-Yu Kou, "We Hold Ourselves Accountable: A Relational View of Team Accountability," *Journal of Business Ethics* 183, no. 3 (November 18, 2023): 691–712, https://doi.org/10.1007/s10551-021-04969-z.

[272] *Organizational Transformation: A Framework for Assessing and Improving Enterprise Architecture Management*, version 2.0 (U.S. Government Accountability Office, August 5, 2010), www.gao.gov/products/gao-10-846g.

[273] Nampuraja Enose Kamalabai, Ilkka Donoghue, and Lea Hannola, "Sustainable Enterprise Architecture: A Critical Imperative for Substantiating Artificial Intelligence," in *2024 Portland International Conference on Management of Engineering and Technology* (IEEE, 2024), https://doi.org/10.23919/PICMET64035.2024.10653061.

[274] Scott Rose, Oliver Borchert, Stu Mitchell, and Sean Connelly, *Zero Trust Architecture* (NIST, August 2020), https://doi.org/10.6028/NIST.SP.800-207.

[275] "Build a Secure Zero Trust Foundation for AI," Microsoft, accessed October 25, 2025, www.microsoft.com/en-us/microsoft-365/business-insights-ideas/resources/build-a-secure-zero-trust-secure-foundation-for-ai.

[276] Lance Hayden, "Designing Common Control Frameworks: A Model for Evaluating Information Technology Governance, Risk, and Compliance Control Rationalization Strategies," *Information Security Journal: A Global Perspective* 18 (2009): 297–305, https://doi.org/10.1080/19393550903324936.

[277] "National Checklist Program," NIST, accessed August 2025, https://ncp.nist.gov/repository.

[278] "What Is the Principle of Least Privilege?," Palo Alto Networks, accessed September 2025, www.paloaltonetworks.com/cyberpedia/what-is-the-principle-of-least-privilege.

[279] Ty Adcock, "The Power of Least Functionality in System Design: A Path to Secure, Efficient, and Resilient Systems," TrainingTraining.Training, January 1, 2025, www.trainingtraining.training/blog/power-of-least-functionality-in-system-design.

[280] Miloslava Plachkinova and Kenneth Knapp, "Least Privilege Across People, Process, and Technology: Endpoint Security Framework," *Journal of Computer Information Systems* 63 (2022): 1153–1165, https://doi.org/10.1080/08874417.2022.2128937.

[281] Isaac Madan, “Securing AI with Least Privilege,” Nightfall, April 3, 2024, www.nightfall.ai/blog/securing-ai-with-least-privilege.

[282] “A Guide to the Data Protection Principles,” UK Information Commissioner’s Office, accessed September 2025, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/.

[283] Prakhar Ganesh, Cuong Tran, Reza Shokri, and Ferdinando Fioretto, “The Data Minimization Principle in Machine Learning,” in *FAccT ’25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2025), 3075–3093, https://doi.org/10.1145/3715275.3732195.

[284] Abigail Goldsteen, Gilad Ezov, Ron Shmelkin, Micha Moffie, and Ariel Farkash, “Data Minimization for GDPR Compliance in Machine Learning Models,” *AI and Ethics* 2 (2022): 477–491, https://doi.org/10.1007/s43681-021-00095-8.

[285] Devansh Saxena, Ji-Youn Jung, Jodi Forlizzi, Kenneth Holstein, and John Zimmerman, “AI Mismatches: Identifying Potential Algorithmic Harms Before AI Development,” in *CHI ’25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2025), https://doi.org/10.1145/3706598.3714098.

[286] Md Shamsujjoha, Qinghua Lu, Dehai Zhao, and Liming Zhu, “Towards AI-Safety-by-Design: A Taxonomy of Runtime Guardrails in Foundation Model Based Systems,” arXiv preprint arXiv:2408.02205 (2024), https://doi.org/10.48550/arXiv.2408.02205.

[287] Tara Capel and Margot Brereton, “What Is Human-Centered About Human-Centered AI? A Map of the Research Landscape,” in *CHI ’23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2023), https://doi.org/10.1145/3544548.3580959.

[288] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran et al.,” Power to the People? Opportunities and Challenges for Participatory AI,” in *EAAMO ’22: Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Association for Computing Machinery, 2022), https://doi.org/10.1145/3551624.3555290.

[289] Conrad Sanderson, David Douglas, Qinghua Lu, Emma Schleiger, Jon Whittle, and Justine Lacey, “AI Ethics Principles in Practice: Perspectives of Designers and Developers," *IEEE Transactions on Technology and Society* 4, no. 2 (June 2023): 171–187, https://doi.org/10.1109/TTS.2023.3257303.

[290] Bo Li, Peng Qi, Bo Liu et al., “Trustworthy AI: From Principles to Practices,” *ACM Computing Survey* 55, no. 9 (September 2023): 1–466, https://doi.org/10.1145/3555803.

[291] Steven Umbrello and Ibo van de Poel, “Mapping Value Sensitive Design onto AI for Social Good Principles,” *AI and Ethics* 1 (2021): 283–296, https://doi.org/10.1007/s43681-021-00038-3.

[292] Lorrie Faith Cranor, "A Framework for Reasoning About the Human in the Loop," in *UPSEC '08: Proceedings of the 1st Conference on Usability, Psychology, and Security* (USENIX Association, 2008), https://dl.acm.org/doi/10.5555/1387649.1387650.

[293] Sriraam Natarajan, Saurabh Mathur, Sahil Sidheekh, Wolfgang Stammer, Kristian Kersting, "Human-in-the-Loop or AI-in-the-Loop? Automate or Collaborate?," *Proceedings of the AAAI Conference on Artificial Intelligence* 39, no. 27 (2025): 28594-28600, https://doi.org/10.1609/aaai.v39i27.35083.

[294] Nicholas Conlon, Nisar R. Ahmed, and Daniel Szafir, "A Survey of Algorithmic Methods for Competency Self-Assessments in Human-Autonomy Teaming," *ACM Computing Surveys* 56, no. 7 (July 2024): 1–31, https://doi.org/10.1145/3616010.

[295] Stephan J. Lemmer, Anhong Guo, and Jason J Corso, "Human-Centered Deferred Inference: Measuring User Interactions and Setting Deferral Criteria for Human-AI Teams," *IUI '23: Proceedings of the 28th International Conference on Intelligent User Interfaces* (Association for Computing Machinery, 2023), 681–694, https://doi.org/10.1145/3581641.3584092.

[296] Vanshika Vats, Marzia Binta Nizam, Minghao Liu et al., "A Survey on Human-AI Teaming with Large Pre-trained Models," arXiv preprint arXiv:2403.04931 (2025), https://doi.org/10.48550/arXiv.2403.04931.

[297] Dian Chen, Han Jun Yoon, Zelin Wan et al., "Advancing Human-Machine Teaming: Concepts, Challenges, and Applications," arXiv preprint arXiv:2503.16518 (2025), https://doi.org/10.48550/arXiv.2503.16518.

[298] *A Guide to Privacy by Design* (Agencia Española de Protección de Datos, October 2019), https://www.aepd.es/guides/guide-to-privacy-by-design.pdf.

[299] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin, "When Machine Learning Meets Privacy: A Survey and Outlook," *ACM Computing Surveys* 54, no. 2 (Association for Computing Machinery, March 2022), https://doi.org/10.1145/3436755.

[300] *AI, Data Governance and Privacy: Synergies and Areas of International Co-Operation* (OECD, June 2024), www.oecd.org/en/publications/ai-data-governance-and-privacy_2476b1a4-en.html.

[301] *Secure by Design: Shifting the Balance of Cybersecurity Risk* (CISA, October 2025), www.cisa.gov/resources-tools/resources/secure-by-design.

[302] "Machine Learning Principles," UK National Cyber Security Centre, May 22, 2024, www.ncsc.gov.uk/collection/machine-learning-principles.

[303] Nate Lee and Laura Voicu, *Securing LLM Backed Systems: Essential Authorization Practices* (Cloud Security Alliance, 2024), https://cloudsecurityalliance.org/artifacts/securing-llm-backed-systems-essential-authorization-practices.

[304] Nate Lee and Ken Huang, *Secure Agentic System Design: A Trait-Based Approach* (Cloud Security Alliance, 2025), https://cloudsecurityalliance.org/artifacts/secure-agentic-system-design.

[305] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, Yi-Chieh Lee, "The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships," in *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2025), https://doi.org/10.1145/3706598.3713429.

[306] Arianna Manzini, Geoff Keeling, Lize Alberts, Shannon Vallor, Meredith Ringel Morris, and Iason Gabriel, "The Code That Binds Us: Navigating the Appropriateness of Human-AI Assistant Relationships," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (2024): 943–957, https://doi.org/10.1609/aies.v7i1.31694.

[307] Rachele Carli, Amro Najjar, and Dena Al-Thani, "Human-Agent Interaction and Human Dependency: Possible New Approaches for Old Challenges," in *HAI '24: Proceedings of the 12th International Conference on Human-Agent Interaction* (Association for Computing Machinery, 2024), 214–223, https://doi.org/10.1145/3687272.3688308.

[308] Jason Phang, Michael Lampe, Lama Ahmad et al., "Investigating Affective Use and Emotional Well-Being on ChatGPT" arXiv preprint arXiv:2504.03888 (2025), https://doi.org/10.48550/arXiv.2504.03888.

[309] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making," *Proceedings of the ACM on Human-Computer Interaction* 5 (April 2021), https://doi.org/10.1145/3449287.

[310] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna, "Explanations Can Reduce Overreliance on AI Systems During Decision-Making," *Proceedings of the ACM on Human-Computer Interaction* 7 (April 2023), https://doi.org/10.1145/3579605.

[311] Cathy Mengying Fang, Auren R. Liu, Valdemar Danry et al., "How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study," arXiv preprint arXiv:2503.17473 (2025), https://doi.org/10.48550/arXiv.2503.17473.

[312] Vagner Figueredo de Santana, Sara E Berger, Heloisa Candello et al., "Responsible Prompting Recommendation: Fostering Responsible AI Practices in Prompting-Time," in *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2025), https://doi.org/10.1145/3706598.3713365.

[313] Zsuzsa Varvasovszky and Ruairí Brugha, "A Stakeholder Analysis," *Health Policy and Planning* 15, no. 3 (September 2000): 338–345, https://doi.org/10.1093/heapol/15.3.338.

[314] Kammi Schmeer, "Stakeholder Analysis Guidelines," in *Policy Toolkit for Strengthening Health Sector Reform*, vol. 1 (LAC HSR, 1999), https://openlmis.org/wp-content/uploads/2018/04/33.pdf.

[315] Advait Deshpande and Helen Sharp, "Responsible AI Systems: Who Are the Stakeholders?," in *AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery, 2022), 227–236, https://doi.org/10.1145/3514094.3534187.

[316] Aubrey L. Mendelow, "Environmental Scanning—The Impact of the Stakeholder Concept," *ICIS 1981 Proceedings* (1981), https://aisel.aisnet.org/icis1981/20.

[317] Ronald K. Mitchell, Bradley R. Agle, and Donna J. Wood, "Toward a Theory of Stakeholder Identification and Salience: Defining the Principle of Who and What Really Counts," *The Academy of Management Review* 22, no. 4 (1997): 853–886, https://doi.org/10.2307/259247.

[318] "Mapping Your Supply Chain," UK National Cyber Security Centre, February 16, 2023, www.ncsc.gov.uk/guidance/mapping-your-supply-chain.

[319] Bart L. MacCarthy, Wafaa A. H. Ahmed, and Guven Demirel, "Mapping the Supply Chain: Why, What and How?," *International Journal of Production Economics* 250 (2022), https://doi.org/10.1016/j.ijpe.2022.108688.

[320] John T. Gardner and Martha C Cooper, "Strategic Supply Chain Mapping Approaches," *Journal of Business Logistics* 24, no. 2 (2003): 37–64, https://doi.org/10.1002/j.2158-1592.2003.tb00045.x.

[321] "Managing Supply Chain Risk to Machine Learning Systems," U.S. National Counterintelligence and Security Center, 2022, www.dni.gov/index.php/ncsc-what-we-do/ncsc-supply-chain-threats.

[322] Justin Sherman, "Securing Data in the AI Supply Chain," Atlantic Council, September 5, 2025, www.atlanticcouncil.org/in-depth-research-reports/issue-brief/securing-data-in-the-ai-supply-chain/.

[323] Jon Boyens, Angela Smith, Nadya Bartol, Kris Winkler, Alex Holbrook, and Matthew Fallon, *Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations* (NIST, May 2022), https://doi.org/10.6028/NIST.SP.800-161r1-upd1.

[324] Iain Barclay and Will Abramson, "Identifying Roles, Requirements and Responsibilities in Trustworthy AI Systems," in *UbiComp/ISWC '21 Adjunct: Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers* (Association for Computing Machinery, 2021), 264–271, https://doi.org/10.1145/3460418.3479344.

[325] Rohan Sharma, "Strategic Insights on the Reporting Structures of AI Executives," in *AI and the Boardroom* (Apress, 2024), https://doi.org/10.1007/979-8-8688-0796-1_27.

[326] Nick Hamilton, Ken Huang, and Michael Roza, *AI Organizational Responsibilities: Governance, Risk Management, Compliance, and Cultural Aspects* (Cloud Security Alliance, 2024), https://cloudsecurityalliance.org/artifacts/ai-organizational-responsibilities-governance-risk-management-compliance-and-cultural-aspects.

[327] Jerry Huang and Ken Huang, *AI Organizational Responsibilities: Core Security Responsibilities* (Cloud Security Alliance, 2024), https://cloudsecurityalliance.org/artifacts/ai-organizational-responsibilities-core-security-responsibilities.

[328] Seung-Bum Yang and Sang Ok Choi, "Employee Empowerment and Team Performance: Autonomy, Responsibility, Information, and Creativity," *Team Performance Management: An International Journal* 15, no. 5–6 (August 21, 2009): 289–301, https://doi.org/10.1108/13527590910983549.

[329] Paul S. Hempel, Zhi-Xue Zhang, and Yulan Han, "Team Empowerment and the Organizational Context: Decentralization and the Contrasting Effects of Formalization," *Journal of Management* 38, no. 2 (2009): 475–501, https://doi.org/10.1177/0149206309342891.

[330] "Workflow Approval Process: How to Create and Build," Solvexia, October 15, 2024, www.solvexia.com/blog/workflow-approval-process-how-to-create-and-build.

[331] Kent Sokoloff, Hadassah Drukarch, Sez Harmon, and Patrick McAndrew, *AI Inventories: Practical Challenges for Organizational Risk Management* (Responsible AI Institute, February 2025), www.responsible.ai/chevron-and-responsible-ai-institute-release-guide-on-ai-inventories-and-risk-management/.

[332] "CIS Inventory Tracking Spreadsheets for v8.1 and v7.1 of the CIS Controls," Center for Internet Security, July 10, 2019, www.cisecurity.org/insights/white-papers/cis-controls-inventory-tracking-spreadsheets.

[333] Samuel Idowu, Daniel Strüber, and Thorsten Berger, "Asset Management in Machine Learning: A Survey," in *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice* (IEEE, 2021), 51–60, https://doi.org/10.1109/ICSE-SEIP52600.2021.00014.

[334] Shaked Rotlevi, "What Is Data Flow Mapping?," Wiz, November 22, 2024, www.wiz.io/academy/data-flow-mapping.

[335] Kabir Barday, Lois Denise Farnsworth, and Toby Spry, "Data Mapping: How to Do It & Why It Matters," video, IAPP, December 15, 2016, https://iapp.org/resources/article/web-conference-data-mapping-how-to-do-it-why-it-matters.

[336] "What Is a Data Flow Diagram (DFD)?," IBM, accessed September 2025, www.ibm.com/think/topics/data-flow-diagram.

[337] Erik Bergström, Fredrik Karlsson, and Rose-Mharie Åhlfeldt, "Developing an Information Classification Method," *Information and Computer Security* 29, no. 2 (August 3, 2021): 209–239, https://doi.org/10.1108/ICS-07-2020-0110.

[338] Farhad Foroughi, "Information Asset Valuation Method for Information Technology Security Risk Assessment," in *Proceedings of the World Congress on Engineering*, vol. 1 (International Association of Engineers, 2008), www.iaeng.org/publication/WCE2008/.

[339] Youngja Park, Wilfried Teiken, Josyula R. Rao, and S. N. Chari, "Data Classification and Sensitivity Estimation for Critical Asset Discovery," *IBM Journal of Research and Development* 60, no. 4,(July 2016): 1–12, https://doi.org/10.1147/JRD.2016.2557638.

[340] Shemlse Gebremedhin Kassa, "IT Asset Valuation, Risk Assessment and Control Implementation Model," ISACA, May 1, 2017, www.isaca.org/resources/isaca-journal/issues/2017/volume-3/it-asset-valuation-risk-assessment-and-control-implementation-model.

[341] Shayne Longpre, Robert Mahari, Anthony Chen et al., "A Large-Scale Audit of Dataset Licensing and Attribution in AI," *Nature Machine Intelligence* 6 (2024): 975–987, www.nature.com/articles/s42256-024-00878-8.

[342] Shayne Longpre, Robert Mahari, Naana Obeng-Marnu et al. "Position: Data Authenticity, Consent, & Provenance for AI Are All Broken: What Will It Take to Fix Them?," *Proceedings of Machine Learning Research* 235 (2024): 32711–32725, https://proceedings.mlr.press/v235/longpre24b.html.

[343] Tomas Vancisin, Loraine Clarke, Mary Orr, and Uta Hinrichs, "Provenance Visualization: Tracing People, Processes, and Practices Through a Data-Driven Approach to Provenance," *Digital Scholarship in the Humanities* 38, no. 3 (2023): 1322–1339, https://academic.oup.com/dsh/article/38/3/1322/7140400.

[344] Boquan Li, Zirui Fu, Mengdi Zhang, Peixin Zhang, Jun Sun, and Xingmei Wang, "Efficient and Universal Watermarking for LLM-Generated Code Detection," arXiv preprint arXiv:2402.07518 (2024), https://doi.org/10.48550/arXiv.2402.07518.

[345] Michael A. Reynolds, Samantha L. Ortiz, Nathaniel J. Jayes, and Olivia M. Carter, "Model Lineage Tracking in Government AI Projects," ResearchGate (August 2025), www.researchgate.net/publication/395415923_Model_Lineage_Tracking_in_Government_AI_Projects.

[346] Alan Chan, Carson Ezell, Max Kaufmann et al., "Visibility into AI Agents," in *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2024), 958–973. https://doi.org/10.1145/3630106.3658948.

[347] "Hardware Bill of Materials (HBOM) Framework for Supply Chain Risk Management," CISA, September 25, 2023, www.cisa.gov/resources-tools/resources/hardware-bill-materials-hbom-framework-supply-chain-risk-management.

[348] "2025 Minimum Elements for a Software Bill of Materials (SBOM)," CISA, August 22, 2025, www.cisa.gov/resources-tools/resources/2025-minimum-elements-software-bill-materials-sbom.

[349] "A Shared Vision of Software Bill of Materials (SBOM) for Cybersecurity," CISA, September 3, 2025, www.cisa.gov/resources-tools/resources/shared-vision-software-bill-materials-sbom-cybersecurity.

[350] "AI Bill of Materials (AIBOM) Project," OWASP Foundation, accessed September 2025, https://owasp.org/www-project-aibom/.

[351] Margaret Mitchell, Simone Wu, Andrew Zaldivar et al., "Model Cards for Model Reporting," in *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2019), 220–229, https://doi.org/10.1145/3287560.3287596.

[352] Isabelle Hupont, David Fernández-Llorca, Sandra Baldassarri, and Emilia Gómez, "Use Case Cards: A Use Case Reporting Framework Inspired by the European AI Act," *Ethics and Information Technology* 26, no. 19 (2024), https://doi.org/10.1007/s10676-024-09757-7.

[353] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury, "Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices," *Proceedings of the ACM on Human-Computer Interaction* 5, no. 1 (April 2021), https://doi.org/10.1145/3449081.

[354] James Ryseff, Brandon F. De Bruhl, and Sydne J. Newberry, *The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed: Avoiding the Anti-Patterns of AI* (RAND, August 13, 2024), www.rand.org/pubs/research_reports/RRA2680-1.html.

[355] Jens Westenberger, Kajetan Schuler, and Dennis Schlegel, "Failure of AI Projects: Understanding the Critical Factors," *Procedia Computer Science* 196 (2022): 69–76, https://doi.org/10.1016/j.procs.2021.11.074.

[356] *Five AI Fails (and How We Can Learn from Them)* (MITRE, August 2021), www.mitre.org/news-insights/publication/five-ai-fails-and-how-we-can-learn-from-them.

[357] *ENISA Threat Landscape 2025* (European Union Agency for Cybersecurity, October 1, 2025), www.enisa.europa.eu/publications/enisa-threat-landscape-2025.

[358] "Cybersecurity Alerts & Advisories," CISA, accessed September 2025, www.cisa.gov/news-events/cybersecurity-advisories.

[359] Government Security Group, "Sourcing a Threat Assessment," UK Government Security, October 3, 2025, www.security.gov.uk/policy-and-guidance/secure-by-design/activities/sourcing-a-threat-assessment/.

[360] Deborah Bodeau, Jennifer Fabius, and Richard Graubart, "How Do You Assess Your Organization's Cyber Threat Level?," MITRE, August 1, 2010, www.mitre.org/news-insights/publication/how-do-you-assess-your-organizations-cyber-threat-level.

[361] Scott Ainslie, Dean Thompson, Sean Maynard, and Atif Ahmad, "Cyber-Threat Intelligence for Security Decision-Making: A Review and Research Agenda for Practice," *Computers & Security* 132 (2023), https://doi.org/10.1016/j.cose.2023.103352.

[362] "CMS Threat Modeling Handbook," CMS CyberGeek, U.S. Centers for Medicare & Medicaid Services, last modified February 21, 2024, https://security.cms.gov/learn/cms-threat-modeling-handbook.

[363] Deborah Bodeau, Catherine McCollum, and David Fox, "Cyber Threat Modeling: Survey, Assessment, and Representative Framework," MITRE, November 6, 2018, www.mitre.org/news-insights/publication/cyber-threat-modeling-survey-assessment-and-representative-framework.

[364] Nataliya Shevchenko, Timothy A. Chick, Paige O'Riordan, Tom Scanlon, and Carol Woody, "Threat Modeling: A Summary of Available Methods," Software Engineering Institute, Carnegie Mellon University, August 9, 2018, www.sei.cmu.edu/library/threat-modeling-a-summary-of-available-methods/.

[365] Xiong Wenjun and Robert Lagerström, "Threat Modeling—A Systematic Literature Review," *Computers & Security* 84 (2019): 53–69, https://doi.org/10.1016/j.cose.2019.03.010.

[366] Nathan VanHoudnos, Carol Smith, Matthew Churilla, Shing-hon Lau, Lauren McIlvenny, and Greg Touhill, "Counter AI: What Is It and What Can You Do About It?," Software Engineering Institute, Carnegie Mellon University, October 7, 2024, www.sei.cmu.edu/library/counter-ai-what-is-it-and-what-can-you-do-about-it/.

[367] Maria Rigaki and Sebastian Garcia, "A Survey of Privacy Attacks in Machine Learning," *ACM Computing Surveys* 56, no. 4 (April 2024), https://doi.org/10.1145/3624010.

[368] Nektaria Kaloudi and Jingyue Li, "The AI-Based Cyber Threat Landscape: A Survey," *ACM Computing Surveys* 53, no. 1 (January 2021), https://doi.org/10.1145/3372823.

[369] Marvin Carr, Suresh Konda, Ira Monarch, Clay F. Walker, and F. Carol Ulrich, "Taxonomy-Based Risk Identification," Software Engineering Institute, Carnegie Mellon University, June 1, 1993, www.sei.cmu.edu/library/taxonomy-based-risk-identification/.

[370] Laura Weidinger, Jonathan Uesato, Maribeth Rauh et al., "Taxonomy of Risks Posed by Language Models," in *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2022), 214–229, https://doi.org/10.1145/3531146.3533088.

[371] Harini Suresh and John Guttag, "A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle," in *EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization* (Association for Computing Machinery, 2021), 1-9, https://doi.org/10.1145/3465416.3483305.

[372] Paul Ward, Jeff Stanley, Ron Ferguson, Colin M. Gladding, and Kevin J. Burns, *Risk Discovery Protocol for AI Assurance* (MITRE, October 2024), www.mitre.org/news-insights/publication/risk-discovery-protocol-ai-assurance-v10.

[373] Steph Batalis, *Anticipating Biological Risk: A Toolkit for Strategic Biosecurity Policy* (CSET, December 2024), https://cset.georgetown.edu/publication/anticipating-biological-risk-a-toolkit-for-strategic-biosecurity-policy/.

[374] *Department of Homeland Security Report on Reducing the Risks at the Intersection of Artificial Intelligence and Chemical, Biological, Radiological, and Nuclear Threats* (DHS, April 26, 2024),

www.dhs.gov/publication/fact-sheet-and-report-dhs-advances-efforts-reduce-risks-intersection-artificial.

[375] Andrew Lohn and Krystal Jackson, *Will AI Make Cyber Swords or Shields?* (CSET, August 2022), https://doi.org/10.51593/2022CA002.

[376] Joint Task Force Transformation Initiative, *Guide for Conducting Risk Assessments* (NIST, September 2012), https://doi.org/10.6028/NIST.SP.800-30r1.

[377] Technical Department of ENISA, Section Risk Management, *Risk Management: Implementation Principles and Inventories for Risk Management/Risk Assessment Methods and Tools* (ENISA, June 2006), www.enisa.europa.eu/publications/risk-management-principles-and-inventories-for-risk-management-risk-assessment-methods-and-tools.

[378] NIST, "NIST Launches ARIA, a New Program to Advance Sociotechnical Testing and Evaluation for AI," news release, May 28, 2024, www.nist.gov/news-events/news/2024/05/nist-launches-aria-new-program-advance-sociotechnical-testing-and.

[379] Leonie Koessler and Jonas Schuett, "Risk Assessment at AGI Companies: A Review of Popular Risk Assessment Techniques from Other Safety-Critical Industries," arXiv preprint arXiv:2307.08823 (2023), https://doi.org/10.48550/arXiv.2307.08823.

[380] Volkan Ervin, "Risk Assessment and Analysis Methods: Qualitative and Quantitative," *ISACA Journal* 2 (2021), www.isaca.org/resources/isaca-journal/issues/2021/volume-2/risk-assessment-and-analysis-methods.

[381] Piorkowski, David, Michael Hind, and John Richards, "Quantitative AI Risk Assessments: Opportunities and Challenges," *Seton Hall Journal of Legislation and Public Policy* 49, no. 3 (2025), https://doi.org/10.60095/HZAV4417.

[382] M. Granger Morgan, Max Henrion, Samuel C. Morris and Deborah A. L. Amaral, "Uncertainty in Risk Assessment," *Environmental Science & Technology* 19, no. 8 (1985): 662–667, https://doi.org/10.1021/es00138a002.

[383] Jun Long and Baruch Fischhoff, "Setting Risk Priorities: A Formal Model," *Risk Analysis* 2, no. 3 (June 2000): 339–351, https://doi.org/10.1111/0272-4332.203033.

[384] Celia Paulsen, Jon Boyens, Nadya Bartol, and Kris Winkler, *Criticality Analysis Process Model: Prioritizing Systems and Components* (NIST, April 2018), https://doi.org/10.6028/NIST.IR.8179.

[385] "A Complete Guide to Risk Prioritization Matrix: Best Practices, Tools, and More," Six Sigma, July 23, 2024, www.6sigma.us/six-sigma-in-focus/risk-prioritization-matrix/.

[386] "The General-Purpose AI Code of Practice," European Commission, accessed August 2025, https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

[387] Gregory Smith, Karlyn D. Stanley, Krystyna Marcinek, Paul Cormarie, and Salil Gunashekar, "Liability for Harms from AI Systems," RAND, November 20, 2024, www.rand.org/pubs/research_reports/RRA3243-4.html.

[388] European Parliamentary Research Service, *The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence* (European Parliament, June 2020), www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)641530.

[389] "Resetting Antidiscrimination Law in the Age of AI," *Harvard Law Review* 138, no. 6 (April 2025), https://harvardlawreview.org/print/vol-138/resetting-antidiscrimination-law-in-the-age-of-ai/.

[390] Civil Rights Division, "Artificial Intelligence and Civil Rights," U.S. Department of Justice Archives, last modified February 3, 2025, www.justice.gov/archives/crt/ai.

[391] UN Office of the High Commissioner on Human Rights, *The Corporate Responsibility to Respect Human Rights: An Interpretive Guide* (UN, 2012), www.ohchr.org/en/publications/special-issue-publications/corporate-responsibility-respect-human-rights-interpretive.

[392] Irene Solaiman, Zeerak Talat, William Agnew et al., "Evaluating the Social Impact of Generative AI Systems in Systems and Society," arXiv preprint arXiv:2306.0594 (2023), https://doi.org/10.48550/arXiv.2306.05949.

[393] Alan F. Winfield, Katina Michael, Jeremy Pitt, and Vanessa Evers, "Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems," *Proceedings of the IEEE* 107, no. 3 (March 2019): 509–517, https://doi.org/10.1109/JPROC.2019.2900622.

[394] National Counterintelligence Security Center, *Protecting Critical Supply Chains: A Guide to Securing Your Supply Chain Ecosystem* (U.S. Office of the Director of National Intelligence, 2024), www.dni.gov/index.php/ncsc-what-we-do/ncsc-supply-chain-threats.

[395] Faisal Aqlan and Sarah S. Lam, "A Fuzzy-Based Integrated Framework for Supply Chain Risk Assessment," *International Journal of Production Economics* 161 (2015): 54–63, https://doi.org/10.1016/j.ijpe.2014.11.013.

[396] Thi Huong Tran, Mario Dobrovnik, and Sebastian Kummer, "Supply Chain Risk Assessment: A Content Analysis-Based Literature Review," *International Journal of Logistics Systems and Management* 31, no. 4 (2018): 562–591, https://doi.org/10.1504/IJLSM.2018.096088.

[397] "OWASP Top 10 Risks for Open Source Software," OWASP Foundation, accessed September 2025, https://owasp.org/www-project-open-source-software-top-10/.

[398] Anja M. Maier, James Moultrie, and P. John Clarkson, "Assessing Organizational Capabilities: Reviewing and Guiding the Development of Maturity Grids," *IEEE Transactions on Engineering Management* 59, no. 1 (2012): 138–159, https://doi.org/10.1109/TEM.2010.2077289.

[399] *Measuring Cybersecurity Workforce Capabilities: Defining a Proficiency Scale for the NICE Framework* (NIST, 2021), www.nist.gov/document/nist-nice-framework-measuring-cybersecurity-workforce-capabilities.

[400] "Cybersecurity Skills and Workforce Frameworks," NIST, last modified September 9, 2025, www.nist.gov/itl/applied-cybersecurity/nice/nice-framework-resource-center/resources/cybersecurity-skills-and.

[401] Alan Turing Institute, *AI Skills for Business Competency Framework*, version 2 (Innovate UK and BridgeAI, January 29, 2024), https://doi.org/10.5281/zenodo.11092677.

[402] Christopher Alberts and David McIntire, "A Systematic Approach for Assessing Workforce Readiness," Software Engineering Institute, Carnegie Mellon University, August 18, 2014, https://sei.cmu.edu/library/a-systematic-approach-for-assessing-workforce-readiness/.

[403] Mesh Flinders and Ian Smalley, "What Is AI Infrastructure?," IBM, accessed October 9, 2025, www.ibm.com/think/topics/ai-infrastructure.

[404] Cristina Silvano, Daniele Ielmini, Fabrizio Ferrandi et al., "A Survey on Deep Learning Hardware Accelerators for Heterogeneous HPC Platforms," *ACM Computing Surveys* 57, no. 11 (November 2025), https://doi.org/10.1145/3729215.

[405] Kevin Petrie and Shawn Rogers, *Preparing and Delivering Data for AI: Adoption Trends, Requirements and Best Practices* (BARC, 2025), https://barc.com/research/data-ai-adoption-trends-requirements-practices/.

[406] Carolina Fortuna, Din Mušić, Gregor Cerar, Andrej Čampa, Panagiotis Kapsalis, and Mihael Mohorčič, "On-Premise Artificial Intelligence as a Service for Small and Medium Size Setups," in *Advances in Engineering and Information Science Toward Smart City and Beyond* (Springer, 2023), https://doi.org/10.1007/978-3-031-29301-6_3.

[407] Gui Alvarenga, "Public Cloud vs. Private Cloud," Crowdstrike, April 11, 2023, www.crowdstrike.com/en-us/cybersecurity-101/cloud-security/public-cloud-vs-private-cloud/.

[408] Ashwin Chavan. "Exploring the Synergy of Cloud and On-Premises Systems—A Case for Hybrid Architectures," *Journal of Computer Science and Technology Studies* 5, no. 3 (2023): 122–141, https://doi.org/10.32996/jcsts.2023.5.3.10.

[409] *Cloud Computing: Benefits, Risks, and Recommendations for Information Security* (ENISA, November 2009), www.enisa.europa.eu/publications/cloud-computing-risk-assessment.

[410] Bader Alouffi, Muhammad Hasnain, Abdullah Alharbi, Wael Alosaimi, Hashem Alyami, and Muhammad Ayaz, "A Systematic Literature Review on Cloud Computing Security: Threats and Mitigation Strategies," *IEEE Access* 9 (2021): 57792–57807, https://doi.org/10.1109/ACCESS.2021.3073203.

[411] Amir Shayan Ahmadian, Daniel Strüber, Volker Riediger, and Jan Jürjens, "Supporting Privacy Impact Assessment by Model-Based Privacy Analysis," in *SAC '18: Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (Association for Computing Machinery, 2018), 1467–1474, https://doi.org/10.1145/3167132.3167288.

[412] Privacy Office, *Privacy Impact Assessments* (DHS, June 2010), www.dhs.gov/publication/privacy-impact-assessment-guidance.

[413] "OPC's Guide to the Privacy Impact Assessment Process," Office of the Privacy Commissioner of Canada, last modified March 28, 2025, www.priv.gc.ca/en/privacy-topics/privacy-impact-assessments/gd_exp_202003/.

[414] "Privacy Impact Assessment (PIA)," CNIL, October 18, 2017, www.cnil.fr/en/privacy-impact-assessment-pia.

[415] Samuel Wairimu, Leonardo Horn Iwaya, Lothar Fritsch, and Stefan Lindskog, "On the Evaluation of Privacy Impact Assessment and Privacy Risk Assessment Methodologies: A Systematic Literature Review," *IEEE Access* 12 (2024): 19625–19650, https://doi.org/10.1109/ACCESS.2024.3360864.

[416] Daniel J. Solove, "A Taxonomy of Privacy," *University of Pennsylvania Law Review* 154, no. 3 (2006), https://scholarship.law.upenn.edu/penn_law_review/vol154/iss3/1/.

[417] Sakib Shahriar, Sonal Allana, Seyed Mehdi Hazratifard, and Rozita Dara, "A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle," *IEEE Access* 11 (2023): 61829–61854, https://doi.org/10.1109/ACCESS.2023.3287195.

[418] "AI Data Security," Australian Signals Directorate, May 23, 2025, www.cyber.gov.au/business-government/secure-design/artificial-intelligence/ai-data-security.

[419] John C. Mankins, *Technology Readiness Levels—A White Paper* (NASA, 1995), www.researchgate.net/publication/247705707_Technology_Readiness_Level_-_A_White_Paper.

[420] John C. Mankins, "Technology Readiness Assessments: A Retrospective," *Acta Astronautica* 65, no. 9–10 (2009): 1216–1223, https://doi.org/10.1016/j.actaastro.2009.03.058.

[421] Alexander Lavin, Ciarán M. Gilligan-Lee, Alessya Visnjic et al., "Technology Readiness Levels for Machine Learning Systems," *Nature Communications* 13 (2022), https://doi.org/10.1038/s41467-022-33128-9.

[422] Victoria Uren and John S. Edwards, "Technology Readiness and the Organizational Journey Towards AI Adoption: An Empirical Study," *International Journal of Information Management* 68 (2023), https://doi.org/10.1016/j.ijinfomgt.2022.102588.

[423] Jan Jöhnk, Malte Weißert, and Katrin Wyrtki, "Ready or Not, AI Comes—An Interview Study of Organizational AI Readiness Factors," *Business and Information System Engineering* 63 (2021): 5–20, https://doi.org/10.1007/s12599-020-00676-7.

[424] *Readiness Assessment Methodology: A Tool of the Recommendation on the Ethics of Artificial Intelligence* (UNESCO, 2023), https://doi.org/10.54678/YHAA4429.
[425] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich, "It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-Off in Machine Learning for Public Policy," in *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2022), 248–266, https://doi.org/10.1145/3531146.3533090.

[426] Jonas Wanner, Kai Heinrich, Christian Janiesch, and Patrick Zschech, "How Much AI Do You Require? Decision Factors for Adopting AI Technology," *Proceedings of the International Conference on Information Systems* 10 (2020), https://aisel.aisnet.org/icis2020/implement_adopt/implement_adopt/10.

[427] Jingyang Li and Guoqiang Li, "Triangular Trade-Off Between Robustness, Accuracy, and Fairness in Deep Neural Networks: A Survey," *ACM Computing Surveys* 57, no. 6 (June 2025), https://doi.org/10.1145/3645088.

[428] Jennifer King, Daniel Ho, Arushi Gupta, Victor Wu, and Helen Webley-Brown, "The Privacy-Bias Tradeoff: Data Minimization and Racial Disparity Assessments in U.S. Government," in *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2023), 492–505, https://doi.org/10.1145/3593013.3594015.

[429] Huiqiang Chen, Tianqing Zhu, Tao Zhang, Wanlei Zhou, and Philip S. Yu, "Privacy and Fairness in Federated Learning: On the Perspective of Tradeoff," *ACM Computing Surveys* 56, no. 2 (February 2024), https://doi.org/10.1145/3606017.

[430] Markus Anderljung, Julian Hazell, and Moritz von Knebel, "Protecting Society from AI Misuse: When Are Restrictions on Capabilities Warranted?," *AI and Society* 40 (2025): 3841–3857, https://doi.org/10.1007/s00146-024-02130-8.

[431] *Microsoft Responsible AI Impact Assessment Guide* (Microsoft, June 2022), www.microsoft.com/en-us/ai/tools-practices.

[432] UK Department for Science, Innovation & Technology, "RAI Institute: Artificial Intelligence Impact Assessment (AIIA)," Gov.uk, April 9, 2024, www.gov.uk/ai-assurance-techniques/rai-institute-artificial-intelligence-impact-assessment-aiia.

[433] Bernd Carsten Stahl, Josephina Antoniou, Nitika Bhalla et al., "A Systematic Review of Artificial Intelligence Impact Assessments," *Artificial Intelligence Review* (March 2023): 1–33, https://doi.org/10.1007/s10462-023-10420-8.

[434] Ministry of Infrastructure and Water Management, *AI Impact Assessment: The Tool for a Responsible AI Project* (Government of the Netherlands, December 2024), www.government.nl/documents/publications/2023/03/02/ai-impact-assessment.

[435] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf, *Assembling Accountability: Algorithmic Impact Assessment for the Public Interest* (Data & Society Research Institute, June 2021), https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/.

[436] Paolo Ceravolo, Ernesto Damiani, Maria Elisa D'Amico et al., "HH4AI: A Methodological Framework for AI Human Rights Impact Assessment Under the EUAI ACT," arXiv preprint arXiv:2503.18994 (2025), https://doi.org/10.48550/arXiv.2503.18994.

[437] David Leslie, Christopher Burr, Mhairi Aitken, Michael Katell, Morgan Briggs, and Cami Rincon, *Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems* (Alan Turing Institute, 2022), https://doi.org/10.5281/zenodo.5981676.

[438] Christopher Summerfield, Lisa P. Argyle, Michiel Bakker et al., "The Impact of Advanced AI Systems on Democracy," *Nature Human Behavior* 9 (2025), https://doi.org/10.1038/s41562-025-02309-z.

[439] *ISO/IEC 42005:2025: Information Technology—Artificial Intelligence (AI)—AI System Impact Assessment* (ISO, 2025), www.iso.org/standard/42005.

[440] Edyta Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia, "Co-designing an AI Impact Assessment Report Template with AI Practitioners and AI Compliance Experts," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (2024): 168–180, https://doi.org/10.1609/aies.v7i1.31627.

[441] *Microsoft Responsible AI Impact Assessment Template* (Microsoft, 2022), www.microsoft.com/en-us/ai/tools-practices.

[442] Stefan Buijsman and Herman Veluwenkamp, "Measuring the Right Thing: Justifying Metrics in AI Impact Assessments," arXiv preprint arXiv:2504.05007 (2025), https://doi.org/10.48550/arXiv.2504.05007.

[443] Miriam Stankovich, Erica Behrens, and Julia Burchell, *Toward Meaningful Transparency and Accountability of AI Algorithms in Public Service Delivery* (DAI, August 2023), www.dai.com/uploads/ai-in-public-service.pdf.

[444] Christopher Theisen, Nuthan Munaiah, Mahran Al-Zyoud, Jeffrey C. Carver, Andrew Meneely, and Laurie Williams, "Attack Surface Definitions: A Systematic Literature Review," *Information and Software Technology* 104 (2018): 94–103, https://doi.org/10.1016/j.infsof.2018.07.008.

[445] "ATT&CK," MITRE, accessed October 8, 2025, https://attack.mitre.org/.

[446] "ATLAS Matrix," MITRE, accessed October 8, 2025, https://atlas.mitre.org/matrices/ATLAS.

[447] "Attack Surface Analysis Cheat Sheet," OWASP Cheat Sheet Series, accessed September 2025, https://cheatsheetseries.owasp.org/cheatsheets/Attack_Surface_Analysis_Cheat_Sheet.html.

[448] Douglas Everson and Long Cheng, "A Survey on Network Attack Surface Mapping," *Digital Threats* 5, no. 2 (June 2024): 1–25, https://doi.org/10.1145/3640019.

[449] *OWASP Machine Learning Security Top 10* (OWASP, 2023), https://mltop10.info/.

[450] *OWASP Top 10 for LLM Applications 2025* (OWASP, November 17, 2024), https://genai.owasp.org/llm-top-10/.

[451] *Agentic AI—Threats and Mitigations* (OWASP, February 2025), https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/.

[452] Martin Husák and Lukáš Sadlek, "Attack Surface Management: State of the Art and Operational Challenges," in *2025 IEEE 11th International Conference on Network Softwarization* (IEEE, 2025), 1–6, https://doi.org/10.1109/NetSoft64993.2025.11080588.

[453] Seyedhamed Ghavamnia, Tapti Palit, Shachee Mishra, and Michalis Polychronakis, "Temporal System Call Specialization for Attack Surface Reduction," presentation at 29th USENIX Security Symposium, 2020, www.usenix.org/conference/usenixsecurity20/presentation/ghavamnia.

[454] Farida Ali, "How to Build a Transparent Relationship with Your Suppliers," *Harvard Business Review*, September 24, 2021, https://hbr.org/2021/09/how-to-build-a-transparent-relationship-with-your-suppliers.

[455] Christian F. Durach and José A. D. Machuca, "A Matter of Perspective—The Role of Interpersonal Relationships in Supply Chain Risk Management," *International Journal of Operations & Production Management* (October 18, 2018): 1866–1887, https://doi.org/10.1108/IJOPM-03-2017-0157.

[456] Darrell Rigby, Zach First, and Dunigan O'Keeffe, "How to Create a Stakeholder Strategy," *Harvard Business Review*, May–June 2023, https://hbr.org/2023/05/how-to-create-a-stakeholder-strategy.

[457] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang, "Stakeholder Participation in AI: Beyond Add Diverse Stakeholders and Stir," arXiv preprint arXiv:2111.01122 (2021), https://doi.org/10.48550/arXiv.2111.01122.

[458] Eric Corbett, Remi Denton, and Sheena Erete, "Power and Public Participation in AI," in *EAAMO '23: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Association for Computing Machinery, 2023), https://doi.org/10.1145/3617694.3623228.

[459] Anat Lior, "Private and Academic AI Collaboration: Opportunities and Challenges to Open Science in the US," *Journal of Open Access to Law* 11, no. 2 (2023), https://doi.org/10.63567/8wfpme67.

[460] Markus Perkmann, Valentina Tartari, Maureen McKelvey et al., "Academic Engagement and Commercialisation: A Review of the Literature on University–Industry Relations," *Research Policy* 42, no. 2 (2013): 423–442, https://doi.org/10.1016/j.respol.2012.09.007.

[461] Peter Cihon, Jonas Schuett, and Seth D. Baum, "Corporate Governance of Artificial Intelligence in the Public Interest," *Information* 12, no. 7 (2021): 275, https://doi.org/10.3390/info12070275.

[462] Slava Jankin Mikhaylov, Marc Esteve, and Averill Campion, "Artificial Intelligence for the Public Sector: Opportunities and Challenges of Cross-Sector Collaboration," *Philosophical Transactions of the Royal Society A* 376, no. 2128 (2018), http://doi.org/10.1098/rsta.2017.0357.

[463] *NIST Cybersecurity Supply Chain Risk Management: Due Diligence Assessment Quick-Start Guide* (NIST, October 2024), https://doi.org/10.6028/NIST.SP.1326.ipd.

[464] *Due Diligence Framework* (Australian Department of Foreign Affairs and Trade, July 2024), www.dfat.gov.au/about-us/publications/due-diligence-framework.

[465] Columbia University School of International Public Affairs, *Quantifying the Costs, Benefits and Risks of Due Diligence for Responsible Business Conduct: Framework and Assessment Tool for Companies* (OECD, June 2016), www.oecd.org/content/dam/oecd/en/topics/policy-sub-issues/due-diligence-guidance-for-responsible-business-conduct/Quantifying-the-Cost-Benefits-Risks-of-Due-Diligence-for-RBC.pdf.

[466] HM Revenue and Customs, "Advice on Applying Supply Chain Due Diligence Principles to Assure Your Labour Supply Chains," Gov.uk, May 13, 2021, www.gov.uk/government/publications/use-of-labour-providers/advice-on-applying-supply-chain-due-diligence-principles-to-assure-your-labour-supply-chains.

[467] Chinasa T. Okolo and Marie Tano, "Moving Toward Truly Responsible AI Development in the Global AI Market," Brookings, October 24, 2024, www.brookings.edu/articles/moving-toward-truly-responsible-ai-development-in-the-global-ai-market/.

[468] William Ho, Xiaowei Xu, and Prasanta K. Dey, "Multi-criteria Decision Making Approaches for Supplier Evaluation and Selection: A Literature Review," *European Journal of Operational Research* 202, no. 1 (2010): 16–24, https://doi.org/10.1016/j.ejor.2009.05.009.

[469] Steve Charkoudian and Omer Tene, "Contracting Around AI: Reading the Fine Print," IAPP, November 27, 2024, https://iapp.org/news/a/contracting-around-ai-reading-the-fine-print.

[470] Ellen English, "AI Terms of Use: Key Issues," *Practical Law*, November 2024, www.reuters.com/practical-law-the-journal/transactional/ai-terms-use-key-issues-2024-11-01/.

[471] Lisa R. Lifshitz, "Avoiding AI Agreement Dystopia: Managing Key Risks in AI Licensing Deals," *Business Law Today*, September 2023, www.americanbar.org/groups/business_law/resources/business-law-today/2023-september/avoiding-ai-agreement-dystopia-managing-key-risks-in-ai-licensing-deals/.

[472] Jessica Bishop and Sarah Stothart, "Artificial Intelligence (AI) Agreements Checklist," *Practical Guidance Journal*, February 2, 2025, www.lexisnexis.com/community/insights/legal/practical-guidance-journal/b/pa/posts/artificial-intelligence-ai-agreements-checklist.

[473] Lama Ahmad, Sandhini Agarwal, Michael Lampe, and Pamela Mishkin, "OpenAI's Approach to External Red Teaming for AI Models and Systems," arXiv preprint arXiv:2503.16431 (2025), https://doi.org/10.48550/arXiv.2503.16431.

[474] Ranjit Singh, Borhane Blili-Hamelin, Carol Anderson et al., *Red-Teaming in the Public Interest* (Data & Society Research Institute, February 9, 2025), https://datasociety.net/library/red-teaming-in-the-public-interest/.
[475] "Supply Chain Security Guidance," UK National Cyber Security Centre, January 28, 2018, www.ncsc.gov.uk/collection/supply-chain-security.

[476] Jon Boyens, Christopher Brown, Chelsea Deane et al., *Validating the Integrity of Computing Devices* (NIST, December 2022), https://doi.org/10.6028/NIST.SP.1800-34.

[477] Shamik Chaudhuri, Kingshuk Dasgupta, Isaac Hepworth et al., *Securing the AI Software Supply Chain* (Google, April 2024), https://research.google/pubs/securing-the-ai-software-supply-chain/?utm_source=shadowai.beehiiv.com&utm_medium=referral&utm_campaign=shadow-ai-2-may-2024.

[478] *AI & Data Ethics: Engage in Data-Centric Innovation in a Consistent, Responsible Way*" (Causeit, 2024), www.digitalfluency.guide/guidebook/data-ethics.

[479] Lorrie Faith Cranor, "Necessary but Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice," *Journal on Telecommunications and High Technology Law* 10 (2012): 273, https://api.semanticscholar.org/CorpusID:10691968.

[480] Adam J. Andreotta, Nin Kirkham, and Marco Rizzi, "AI, Big Data, and the Future of Consent," *AI & Society* 37 (2022): 1715–1728, https://doi.org/10.1007/s00146-021-01262-5.

[481] Ben Wolford, "What Is a GDPR Data Processing Agreement?," GDPR.eu, January 15, 2023, https://gdpr.eu/what-is-data-processing-agreement/.

[482] Florian Schaub, Rebecca Balebako, Adam L. Durity, and Lorrie Faith Cranor, "A Design Space for Effective Privacy Notices," in *SOUPS 2015 Proceedings: Symposium on Usable Privacy and Security* (USENIX Association, 2015), www.usenix.org/conference/soups2015/proceedings/presentation/schaub.

[483] Office of Technology and The Division of Privacy and Identity Protection, "AI (and Other) Companies: Quietly Changing Your Terms of Service Could Be Unfair or Deceptive," *Technology Blog*, U.S. Federal Trade Commission, February 13, 2024), www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/02/ai-other-companies-quietly-changing-your-terms-service-could-be-unfair-or-deceptive.

[484] Courtney C. Radsch, "The Case for Consent in the AI Data Gold Rush," Brookings, January 16, 2025, www.brookings.edu/articles/the-case-for-consent-in-the-ai-data-gold-rush/.

[485] Lin Kyi, Amruta Mahuli, M. Six Silberman, Reuben Binns, Jun Zhao, and Asia J. Biega, "Governance of Generative AI in Creative Work: Consent, Credit, Compensation, and Beyond," in *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2025), https://doi.org/10.1145/3706598.3713799.

[486] Yuji Roh, Geon Heo, and Steven Euijong Whang, "A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective," *IEEE Transactions on Knowledge and Data Engineering* 33, no. 4 (April 2021): 328–1347, https://doi.org/10.1109/TKDE.2019.2946162.

[487] Daniel J. Solove and Woodrow Hartzog, "The Great Scrape: The Clash Between Scraping and Privacy," *California Law Review* 113, no. 1521 (2025), https://dx.doi.org/10.2139/ssrn.4884485.

[488] Vlad Krotov and Leigh Johnson, "Big Web Data: Challenges Related to Data, Technology, Legality, and Ethics," *Business Horizons* 66, no. 4 (2023): 481–491, https://doi.org/10.1016/j.bushor.2022.10.001.

[489] Jayasankar Jayachandran and Vijay Arni, "Traversing the Ethical Landscape of Data Scraping for AI," SSRN, December 8, 2023, http://dx.doi.org/10.2139/ssrn.4666354.

[490] Stefan Baack, "A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl," in *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2024), 2199–2208, https://doi.org/10.1145/3630106.3659033.

[491] Won Kyung Jung and Hun Yeong Kwon, "Privacy and Data Protection Regulations for AI Using Publicly Available Data: Clearview AI Case," in *ICEGOV '24: Proceedings of the 17th International Conference on Theory and Practice of Electronic Governance* (Association for Computing Machinery, 2024), 48–55, https://doi.org/10.1145/3680127.3680200.

[492] Pinlong Zhao, Weiyao Zhu, Pengfei Jiao, Di Gao, and Ou Wu, "Data Poisoning in Deep Learning: A Survey," arXiv preprint arXiv:2503.22759 (2025), https://doi.org/10.48550/arXiv.2503.22759.

[493] Nayna Jaen, "Determining What AI Data You Need and How to Source It," RWS, April 8, 2024, www.rws.com/artificial-intelligence/train-ai-data-services/blog/determining-ai-data-needs-and-sourcing/.

[494] *Securing Artificial Intelligence (SAI) Data Supply Chain Security* (ETSI, January 2025), www.etsi.org/committee/technical-committee-tc-securing-artificial-intelligence-sai.

[495] Philippe De Wilde, Payal Arora, Fernando Buarque de Lima Neto et al., *Recommendations on the Use of Synthetic Data to Train AI Models* (UN University, February 14, 2024), https://unu.edu/publication/recommendations-use-synthetic-data-train-ai-models.

[496] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson, "The Curse of Recursion: Training on Generated Data Makes Models Forget," arXiv preprint arXiv:2305.17493 (2023), https://doi.org/10.48550/arXiv.2305.17493.

[497] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey et al., "Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data," arXiv preprint arXiv:2404.01413 (2024), https://doi.org/10.48550/arXiv.2404.01413.

[498] Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah, "How Bad Is Training on Synthetic Data? A Statistical Analysis of Language Model Collapse," arXiv preprint arXiv:2404.05090 (2024), https://doi.org/10.48550/arXiv.2404.05090.

[499] Department for Science, Innovation, and Technology, "Guidelines for AI Procurement," Gov.uk, June 8, 2020, www.gov.uk/government/publications/guidelines-for-ai-procurement/guidelines-for-ai-procurement.

[500] *Adopting AI Responsibly: Guidelines for Procurement of AI Solutions by the Private Sector* (World Economic Forum, June 2023), www.weforum.org/publications/adopting-ai-responsibly-guidelines-for-procurement-of-ai-solutions-by-the-private-sector/.

[501] *LLM03:2025 Supply Chain* (OWASP, 2025), https://genai.owasp.org/llmrisk/llm032025-supply-chain/.

[502] "ML06:2023 ML Supply Chain Attacks," OWASP, accessed September 2025, https://owasp.org/www-project-machine-learning-security-top-10/docs/ML06_2023-AI_Supply_Chain_Attacks.html.

[503] Peerachai Banyongrakkul, Mansooreh Zahedi, Patanamon Thongtanunam, Christoph Treude, and Haoyu Gao, "From Release to Adoption: Challenges in Reusing Pre-trained AI Models for Downstream Developers," arXiv preprint arXiv:2506.23234 (2025), https://doi.org/10.48550/arXiv.2506.23234.

[504] Ben Cottier, Josh You, Natalia Martemianova, and David Owen, "How Far Behind Are Open Models?," Epoch AI, November 4, 2024), https://epoch.ai/blog/open-models-report.

[505] Robert Wolfe, Isaac Slaughter, Bin Han et al., "Laboratory-Scale AI: Open-Weight Models Are Competitive with ChatGPT Even in Low-Resource Settings," in *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2024), 1199–1210. https://doi.org/10.1145/3630106.3658966.

[506] Ernesto Lang Oreamuno, Rina Mrinmoyee Khan, Abdul Ali Bangash, Catherine Stinson, and Bram Adams, "The State of Documentation Practices of Third-Party Machine Learning Models and Datasets," *IEEE Software* 41, no. 5 (2024): 52–59, https://doi.org/10.1109/MS.2024.3366111.

[507] Jiya Manchanda, Laura Boettcher, Matheus Westphalen, and Jasser Jasser, "The Open Source Advantage in Large Language Models (LLMs)," arXiv preprint arXiv:2412.12004 (2025), https://doi.org/10.48550/arXiv.2412.12004.

[508] Dominik Hintersdorf, Lukas Struppek, and Kristian Kersting, "Balancing Transparency and Risk: An Overview of the Security and Privacy Risks of Open-Source Machine Learning Models," in *Bridging the Gap Between AI and Reality* (Springer, 2023), 269–283, https://link.springer.com/chapter/10.1007/978-3-031-73741-1_16.

[509] Hongling Zheng, Li Shen, Anke Tang et al., "Learning from Models Beyond Fine-Tuning," *Nature Machine Intelligence* 7 (2025), https://doi.org/10.1038/s42256-024-00961-0.

[510] Zheda Mai, Arpita Chowdhury, Ping Zhang et al., "Fine-Tuning Is Fine, if Calibrated," *Advances in Neural Information Processing Systems* 37 (2024), https://proceedings.neurips.cc/paper_files/paper/2024/hash/f573c36434796efe066d2f4cf3349e7f-Abstract-Conference.html.

[511] Xiangyu Qi, Yi Zeng, Tinghao Xie et al., "Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!," arXiv preprint arXiv:2310.03693 (2023), https://doi.org/10.48550/arXiv.2310.03693.

[512] Enduring Security Framework Working Group, *Securing the Software Supply Chain: Recommended Practices Guide for Customers* (CISA, October 2022), www.cisa.gov/resources-tools/resources/securing-software-supply-chain-recommended-practices-guide-suppliers-and.

[513] "Discover the Best AI Websites and Tools," Toolify, accessed September 2025, www.toolify.ai/.

[514] "Independent Analysis of AI," Artificial Analysis, accessed August 2025, https://artificialanalysis.ai/.

[515] Artem Vysotsky and Sergey Vysotsky, "Useful Tools to Compare AI Models," Writingmate, March 22, 2025, https://writingmate.ai/blog/useful-tools-to-compare-ai-models.

[516] Francicso Javier Campos Zabala, "Selecting AI Tools and Platforms," in *Grow Your Business with AI* (Apress, 2023), https://doi.org/10.1007/978-1-4842-9669-1_16.

[517] René von Schomberg and Jonathan Hankins, eds., *International Handbook on Responsible Innovation: A Global Resource* (Edward Elgar Publish, 2019), https://doi.org/10.4337/9781784718862.

[518] *Advancing Responsible AI Innovation: A Playbook* (World Economic Forum, September 2025), https://www.weforum.org/publications/advancing-responsible-ai-innovation-a-playbook/.

[519] Department for Science, Innovation, and Technology, "The Model for Responsible Innovation," Gov.uk, November 15, 2024, www.gov.uk/government/publications/the-model-for-responsible-innovation/the-model-for-responsible-innovation.

[520] Matheus de Morais Leça, Mariana Bento, and Ronnie de Souza Santos, "Responsible AI in the Software Industry: A Practitioner-Centered Perspective," arXiv preprint arXiv:2412.07620 (2024), https://doi.org/10.48550/arXiv.2412.07620.

[521] Thomas Stober and Uwe Hansmann, *Agile Software Development: Best Practices for Large Software Development Projects* (Springer, 2010), https://link.springer.com/content/pdf/10.1007/978-3-540-70832-2.pdf.

[522] Lasse Wrobel, Christian Dietzmann, and Rainer Alt, "Ready for Managing AI Projects? An Analysis of AI Project Management Frameworks," in *Proceedings of the 58th Hawaii International Conference on System Sciences* (University of Hawai'i, 2025), https://doi.org/10.24251/HICSS.2025.662.

[523] Beyza Eken, Samodha Pallewatta, Nguyen Tran, Ayse Tosun, and Muhammad Ali Babar, "A Multivocal Review of MLOps Practices, Challenges and Open Issues," *ACM Computing Surveys* 58, no. 2 (January 2026), https://doi.org/10.1145/3747346.

[524] Michael Stone, Chinedum Irrechukwu, Harry Perper, Devin Wynne, and Leah Kauffman, *IT Asset Management* (NIST, September 2018), http://doi.org/10.6028/NIST.SP.1800-5.

[525] Carnegie Mellon University, *CRR Supplemental Resource Guide*, vol. 1, *Asset Management*, version 1.1 (DHS Cyber Security Evaluation Program, 2016), www.cisa.gov/resources-tools/resources/cyber-resilience-review-supplemental-resource-guides.

[526] Carnegie Mellon University, *CRR Supplemental Resource Guide*, vol. 8, *External Dependencies Management*, version 1.1 (DHS Cyber Security Evaluation Program, 2016), www.cisa.gov/resources-tools/resources/cyber-resilience-review-supplemental-resource-guides.

[527] Douglas Robbins, Ozgur Eris, Ariel Kapusta, Lashon Booker, and Paul Ward, *AI Assurance: A Repeatable Process for Assuring AI-Enabled Systems* (MITRE, June 2024), www.mitre.org/news-insights/publication/ai-assurance-repeatable-process-assuring-ai-enabled-systems.

[528] Arnold Johnson, Kelley Dempsey, Ron Ross, Sarbari Gupta, and Dennis Bailey, *Guide for Security-Focused Configuration Management of Information Systems* (NIST, August 2011), https://doi.org/10.6028/NIST.SP.800-128.

[529] Carnegie Mellon University, *CRR Supplemental Resource Guide*, vol. 2, *Controls Management*, version 1.1 (DHS Cyber Security Evaluation Program, 2016), www.cisa.gov/resources-tools/resources/cyber-resilience-review-supplemental-resource-guides.

[530] Carnegie Mellon University, *CRR Supplemental Resource Guide*, vol. 7, *Risk Management*, version 1.1 (DHS Cyber Security Evaluation Program, 2016), www.cisa.gov/resources-tools/resources/cyber-resilience-review-supplemental-resource-guides.

[531] Jon Boyens, Celia Paulsen, Nadya Bartol, Kris Winkler, and James Gimbi, *Key Practices in Cyber Supply Chain Risk Management: Observations from Industry* (NIST, February 2021), https://doi.org/10.6028/NIST.IR.8276.

[532] Carnegie Mellon University, *CRR Supplemental Resource Guide*, vol. 4, *Vulnerability Management*, version 1.1 (DHS Cyber Security Evaluation Program, 2016), www.cisa.gov/resources-tools/resources/cyber-resilience-review-supplemental-resource-guides.

[533] Carnegie Mellon University, *CRR Supplemental Resource Guide*, vol. 3, *Configuration and Change Management*, version 1.1 (DHS Cyber Security Evaluation Program, 2016), www.cisa.gov/resources-tools/resources/cyber-resilience-review-supplemental-resource-guides.

[534] Murugiah Souppaya and Karen Scarfone, *Guide to Enterprise Patch Management Planning: Preventive Maintenance for Technology* (NIST, April 2022), https://doi.org/10.6028/NIST.SP.800-40r4.

[535] Carnegie Mellon University, *CRR Supplemental Resource Guide*, vol. 5, *Incident Management*, version 1.1 (DHS Cyber Security Evaluation Program, 2016), www.cisa.gov/resources-tools/resources/cyber-resilience-review-supplemental-resource-guides.

[536] Carnegie Mellon University, *CRR Supplemental Resource Guide*, vol. 6, *Service Continuity Management*, version 1.1 (DHS Cyber Security Evaluation Program, 2016), www.cisa.gov/resources-tools/resources/cyber-resilience-review-supplemental-resource-guides.

[537] Carnegie Mellon University, *CRR Supplemental Resource Guide*, vol. 9, *Training and Awareness*, version 1.1 (DHS Cyber Security Evaluation Program, 2016), www.cisa.gov/resources-tools/resources/cyber-resilience-review-supplemental-resource-guides.

[538] Carnegie Mellon University, *CRR Supplemental Resource Guide*, vol. 10, *Situational Awareness*, version 1.1 (DHS Cyber Security Evaluation Program, 2016), www.cisa.gov/resources-tools/resources/cyber-resilience-review-supplemental-resource-guides.

[539] David Rock and Heidi Grant, "Why Diverse Teams Are Smarter," *Harvard Business Review* (November 4, 2016), https://hbr.org/2016/11/why-diverse-teams-are-smarter.

[540] Ron Carucci, "One More Time: Why Diversity Leads to Better Team Performance," *Forbes*, January 24, 2024, www.forbes.com/sites/roncarucci/2024/01/24/one-more-time-why-diversity-leads-to-better-team-performance/.

[541] Nichol Bradford, "Why Diversity in AI Makes Better AI for All: The Case for Inclusivity and Innovation," SHRM, October 7, 2024, www.shrm.org/topics-tools/flagships/ai-hi/why-diversity-in-ai-makes-better-ai-for-all--the-case-for-inclus.

[542] *Understanding Talent Scarcity: AI & Equity Report* (Ranstad, November 2024), www.randstad.com/randstad-ai-equity/.

[543] Jeffrey Brown, Tina Park, Jiyoo Chang, Mckane Andrus, Alice Xiang, and Christine Custis, "Attrition of Workers with Minoritized Identities on AI Teams," in *EAAMO '22: Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Association for Computing Machinery, 2022), https://doi.org/10.1145/3551624.3555304.

[544] Elizabeth Mannix and Margaret A. Neale, "What Differences Make a Difference? The Promise and Reality of Diverse Teams in Organizations," *Psychological Science in the Public Interest* 6, no. 2 (2005): 31–55, https://doi.org/10.1111/j.1529-1006.2005.00022.x.

[545] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy, "How AI Developers Overcome Communication Challenges in a Multidisciplinary Team: A Case Study," *Proceedings ACM on Human-Computer Interaction* 5, no. 1 (2021), https://doi.org/10.1145/3449205.

[546] Karen Kent and Murugiah Souppaya, *Guide to Computer Security Log Management* (NIST, September 2006), https://doi.org/10.6028/NIST.SP.800-92.

[547] "Best Practices for Event Logging and Threat Detection," Australian Signals Directorate, August 22, 2024, www.cyber.gov.au/business-government/detecting-responding-to-threats/event-logging/best-practices-event-logging-threat-detection.

[548] "Logging Cheat Sheet," OWASP Cheat Sheet Series, accessed October 2025, https://cheatsheetseries.owasp.org/cheatsheets/Logging_Cheat_Sheet.html.

[549] Piero Bonatti, Sabrina Kirrane, Axel Polleres, and Rigo Wenning, "Transparent Personal Data Processing: The Road Ahead," in *Computer Safety, Reliability, and Security* (Springer, 2017), 337–349, https://doi.org/10.1007/978-3-319-66284-8_28.

[550] Ding Yuan, Soyeon Park, Peng Huang et al., "Be Conservative: Enhancing Failure Diagnosis with Proactive Logging," in *10th USENIX Symposium on Operating Systems Design and Implementation (OSDI '12)* (USENIX Association, 2012), 293–306, www.usenix.org/system/files/conference/osdi12/osdi12-final-109.pdf.

[551] Guoping Rong, Yangchen Xu, Shenghui Gu, He Zhang, and Dong Shao, "Can You Capture Information as You Intend To? A Case Study on Logging Practice in Industry," in *2020 IEEE International Conference on Software Maintenance and Evolution* (IEEE, 2020), 12–22, https://doi.org/10.1109/ICSME46990.2020.00012.

[552] Patrick Loic Foalem, Leuson Da Silva, Foutse Khomh, Heng Li, and Ettore Merlo, "Logging Requirement for Continuous Auditing of Responsible Machine Learning-Based Applications," *Empirical Software Engineering* 30, no. 97 (2025), https://doi.org/10.1007/s10664-025-10656-8.

[553] *ISO/IEC DIS 24970: Artificial Intelligence—AI System Logging* (ISO, forthcoming), www.iso.org/standard/88723.html.

[554] Nazatul Nurlisa Zolkifli, Amir Ngah, and Aziz Deraman, "Version Control System: A Review," *Procedia Computer Science* 135 (2018): 408–415, https://doi.org/10.1016/j.procs.2018.08.191.

[555] Matei Zaharia, Andrew Chen, Aaron Davidson et al., "Accelerating the Machine Learning Lifecycle with MLflow," *IEEE Data Engineering Bulletin* 41, no. 4 (2018): 39–45, http://sites.computer.org/debull/A18dec/p39.pdf.

[556] Maria Priestley, Fionntán O'Donnell, and Elena Simperl, "A Survey of Data Quality Requirements that Matter in ML Development Pipelines," *ACM Journal of Data and Information Quality* 15, no. 2 (2023): 1–39, https://dl.acm.org/doi/10.1145/3592616.

[557] Tianhao Wang, Yi Zeng, Ming Jin, and Ruoxi Jia, "A Unified Framework for Task-Driven Data Quality Management," arXiv preprint arXiv:2106.05484 (2021), https://doi.org/10.48550/arXiv.2106.05484.

[558] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino, "Methodologies for Data Quality Assessment and Improvement," *ACM Computing Surveys* 41, no. 3 (July 2009), https://doi.org/10.1145/1541880.1541883.

[559] Jason Hausenloy, Duncan McClements, and Madhavendra Thakur, "Towards Data Governance of Frontier AI Models," arXiv preprint arXiv:2412.03824 (2025), https://arxiv.org/abs/2412.03824v1.

[560] Iason Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds and Machines* 30 (2020): 411–437, https://doi.org/10.1007/s11023-020-09539-2.

[561] Jiaming Ji, Tianyi Qiu, Boyuan Chen et al., "AI Alignment: A Contemporary Survey," *ACM Computing Surveys* (October 2025), https://doi.org/10.1145/3770749.

[562] Mary Phuong, Roland S. Zimmermann, Ziyue Wang et al., "Evaluating Frontier Models for Stealth and Situational Awareness," arXiv preprint arXiv:2505.01420 (2025), https://doi.org/10.48550/arXiv.2505.01420.

[563] Andrea Tocchetti, Lorenzo Corti, Agathe Balayn et al., "A.I. Robustness: A Human-Centered Perspective on Technological Challenges and Opportunities," *ACM Computing Surveys* 57, no. 6 (June 2025), https://doi.org/10.1145/3665926.

[564] Tim G. J. Rudner and Helen Toner, *Key Concepts in AI Safety: Robustness and Adversarial Examples* (CSET, March 2021). https://doi.org/10.51593/20190041.

[565] Yukun Zhao, Lingyong Yan, Weiwei Sun et al., "Improving the Robustness of Large Language Models via Consistency Alignment," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (ELRA and ICCL, 2024), 8931–8941, https://aclanthology.org/2024.lrec-main.782/.

[566] Youpeng Li, Xinda Wang, Fuxun Yu, Lichao Sun, Wenbin Zhang, and Xuyu Wang, "FedCAP: Robust Federated Learning via Customized Aggregation and Personalization," in *2024 Annual Computer Security Applications Conference* (IEEE, 2024), 747–760, https://doi.org/10.1109/ACSAC63791.2024.00067.

[567] Jie Zhang, Bo Li, Chen Chen et al., "Delving into the Adversarial Robustness of Federated Learning," *Proceedings of the AAAI Conference on Artificial Intelligence* 37, no. 9 (2023): 11245–11253, https://doi.org/10.1609/aaai.v37i9.26331.

[568] Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell, "Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks," in *2023 IEEE Conference on Secure and Trustworthy Machine Learning* (IEEE, 2023), 464–483, https://doi.org/10.1109/SaTML54575.2023.00039.

[569] Q. Vera Liao, Daniel Gruen, and Sarah Miller, "Questioning the AI: Informing Design Practices for Explainable AI User Experiences," in *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2020), 1–15, https://doi.org/10.1145/3313831.3376590.

[570] Dhanesh Ramachandram, Himanshu Joshi, Judy Zhu, Dhari Gandhi, Lucas Hartman, and Ananya Raval, "Transparent AI: The Case for Interpretability and Explainability," arXiv preprint arXiv:2507.23535 (2025), https://doi.org/10.48550/arXiv.2507.23535.

[571] Fadi Al Machot, Martin Thomas Horsch, and Habib Ullah, "Building Trustworthy AI: Transparent AI Systems via Language Models, Ontologies, and Logical Reasoning (TranspNet)," in *Designing the Conceptual Landscape for a XAIR Validation Infrastructure*, eds. Fadi Al Machot, Martin Thomas Horsch, and Sebastian Scholze (Springer, 2025), https://doi.org/10.1007/978-3-031-89274-5_3.

[572] Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin, "A Review of Applications in Federated Learning," *Computers and Industrial Engineering* 149 (2020), https://doi.org/10.1016/j.cie.2020.106854.

[573] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao, "A Survey on Federated Learning," *Knowledge-Based Systems* 216 (2021), https://doi.org/10.1016/j.knosys.2021.106775.

[574] Meng Hao, Hongwei Li, Xizhao Luo, Guowen Xu, Haomiao Yang, and Sen Liu, "Efficient and Privacy-Enhanced Federated Learning for Industrial Artificial Intelligence," *IEEE Transactions on Industrial Informatics* 16, no. 10 (IEEE, 2020): 6532–6542, https://ieeexplore.ieee.org/abstract/document/8859260.

[575] Wenqi Wei and Ling Liu, "Trustworthy Distributed AI Systems: Robustness, Privacy, and Governance," *ACM Computing Surveys* 57, no. 6 (June 2025), https://doi.org/10.1145/3645102.

[576] Lea Demelius, Roman Kern, and Andreas Trügler, "Recent Advances of Differential Privacy in Centralized Deep Learning: A Systematic Survey," *ACM Computing Surveys* 57, no. 6 (June 2025): 1–28, https://doi.org/10.1145/3712000.

[577] Reza Shokri and Vitaly Shmatikov, "Privacy-Preserving Deep Learning," in *CCS '15: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Association for Computing Machinery, 2015), 1310–1321, https://doi.org/10.1145/2810103.2813687.

[578] Cuong Tran and Ferdinando Fioretto, "Data Minimization at Inference Time," *Advances in Neural Information Processing Systems* 36 (2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/e48880ea81caa7836e6a0694049093ae-Abstract-Conference.html.

[579] Murugiah Souppaya, Karen Scarfone, and Donna Dodson, *Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities* (NIST, February 2022), https://doi.org/10.6028/NIST.SP.800-218.

[580] "Guidelines for Secure AI System Development," UK National Cyber Security Centre, November 27, 2023, www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development.

[581] Harold Booth, Murugiah Souppaya, Apostol Vassilev, Michael Ogata, Martin Stanley, and Karen Scarfone, *Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile* (NIST, July 2024), https://doi.org/10.6028/NIST.SP.800-218A.

[582] Frank Li and Vern Paxson, "A Large-Scale Empirical Study of Security Patches," *in CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (Association for Computing Machinery, 2017), 2201–2215, https://doi.org/10.1145/3133956.3134072.

[583] Adam D. G. Jenkins, Linsen Liu, Maria K. Wolters, and Kami Vaniea, "Not as Easy as Just Update: Survey of System Administrators and Patching Behaviours," in *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2024), https://doi.org/10.1145/3613904.3642456.

[584] Nesara Dissanayake, Mansooreh Zahedi, Asangi Jayatilaka, and Muhammad Ali Babar, "Why, How and Where of Delays in Software Security Patch Management: An Empirical Investigation in the Healthcare Sector," *Proceedings of the ACM Human-Computer Interaction* 6, no. 2 (November 2022), https://doi.org/10.1145/3555087.

[585] Florenz A. Martin and William P. Rey, "Patch Perfect: System Administrator Strategies for Effective Patch Management and Securing Systems, Minimizing Risks," in *2024 International Conference on Control, Robotics and Informatics* (IEEE, 2024), 1-6, https://doi.org/10.1109/ICCRI64298.2024.00009.

[586] Nesara Dissanayake, Asangi Jayatilaka, Mansooreh Zahedi, and Muhammad Ali Babar, "An Empirical Study of Automation in Software Security Patch Management," in *ASE '22: Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering* (Association for Computing Machinery, 2023), https://doi.org/10.1145/3551349.3556969.

[587] *Best Practices for Planning and Managing Physical Security Resources: An Interagency Security Committee Guide* (Interagency Security Committee, December 2015), www.cisa.gov/resources-tools/resources/isc-best-practices-planning-and-managing-physical-security-resources.

[588] "Guidelines for Physical Security," Australian Signals Directorate, September 4, 2025, www.cyber.gov.au/business-and-government/cyber-security-frameworks/ism/cybersecurity-guidelines/guidelines-for-physical-security.

[589] *Recommended Practices for Safety and Health Programs* (Occupational Safety and Health Administration, October 2016), www.osha.gov/safety-management/hazard-prevention.

[590] Joint Task Force, *Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy* (NIST, December 2018), https://doi.org/10.6028/NIST.SP.800-37r2.

[591] *Artificial Intelligence Risk Management Framework* (NIST, January 2023), https://doi.org/10.6028/NIST.AI.100-1.

[592] Joint Task Force Transformation Initiative, *Managing Information Security Risk* (NIST, March 2011), https://doi.org/10.6028/NIST.SP.800-39.

[593] Anat Lior, "E/Insuring AI: The Role of Insurance in Artificial Intelligence Regulation," *Harvard Journal of Law & Technology* 35, no. 2 (2022), https://ssrn.com/abstract=4266259.

[594] J. Pedro Mendes, Miguel Marques, and Carlos Guedes Soares, "Risk Avoidance in Strategic Technology Adoption," *Journal of Modelling in Management* 19, no. 5 (October 2024): 1485–1509, https://doi.org/10.1108/JM2-10-2023-0221.

[595] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside, *An Overview of Catastrophic AI Risks* (Center for AI Safety, 2023), https://safe.ai/ai-risk.

[596] Stuart Russell, Charbel-Raphaël Segerie, Niki Iliadis, and Tereza Zoumpalova, "AI Governance Through Global Red Lines Can Help Prevent Unacceptable Risks," OECD, September 22, 2025, https://oecd.ai/en/wonk/ai-governance-through-global-red-lines-can-help-prevent-unacceptable-risks.

[597] Sven Ove Hansson, "Ethical Criteria of Risk Acceptance," *Erkenntnis* 59 (2003): 291–309, https://doi.org/10.1023/A:1026005915919.

[598] "Employment Screening: A Good Practice Guide," UK National Protective Security Authority, last modified November 7, 2025, www.npsa.gov.uk/specialised-guidance/insider-risk-guidance/employment-screening.

[599] "Resources for Onboarding and Employment Screening Fact Sheet," CISA, July 25, 2024, www.cisa.gov/resources-tools/resources/resources-onboarding-and-employment-screening-fact-sheet.

[600] "Insider Threat Mitigation Resources and Tools," CISA, accessed October 8, 2025, www.cisa.gov/topics/physical-security/insider-threat-mitigation/resources-and-tools.

[601] Comptroller General of the United States, *Standards for Internal Control in the Federal Government* (Government Accountability Office, May 2025), www.gao.gov/greenbook.

[602] Jessica Ackerman, Theresa Koursaris, Jim Traeger, and Reshma Shah, *The Private Company Guide to Effective Internal Controls* (Deloitte, 2021), www.deloitte.com/us/en/services/audit-assurance/articles/effective-internal-controls-guide.html.

[603] Stefano Ferroni, "Implementing Segregation of Duties: A Practical Experience Based on Best Practices," ISACA, May 19, 2016, www.isaca.org/resources/isaca-journal/issues/2016/volume-3/implementing-segregation-of-duties-a-practical-experience-based-on-best-practices.

[604] *Managing Risk of Adverse/Involuntary Employee Separations: An Interagency Security Committee Guide* (CISA, 2024), www.cisa.gov/resources-tools/resources/isc-guide-managing-risk-adverseinvoluntary-employee-separations.

[605] David Temoshok, Diana Proud-Madruga, Yee-Yin Choong et al., *Digital Identity Guidelines* (NIST, July 2025), https://doi.org/10.6028/NIST.SP.800-63-4.

[606] *Identity and Access Management: Recommended Best Practices Guide for Administrators* (National Security Agency and CISA, March 2023), www.cisa.gov/news-events/alerts/2023/03/21/cisa-and-nsa-release-enduring-security-framework-guidance-identity-and-access-management.

[607] "Introduction to Identity and Access Management," UK National Cyber Security Centre, January 22, 2022, www.ncsc.gov.uk/guidance/introduction-identity-and-access-management.

[608] Canadian Centre for Cyber Security, "Best Practices for Passphrases and Passwords (ITSAP.30.032)," Government of Canada, February 2024, www.cyber.gc.ca/en/guidance/best-practices-passphrases-and-passwords-itsap30032.

[609] "Require Strong Passwords: Enforcing a Password Manager Protects Your Business," CISA, accessed September 2025, www.cisa.gov/secure-our-world/require-strong-passwords.

[610] *Selecting Secure Multi-Factor Authentication Solutions* (NSA, September 2020), www.nsa.gov/Press-Room/News-Highlights/Article/Article/2356020/nsa-releases-cybersecurity-guidance-selecting-and-safely-using-multifactor-auth/.

[611] *Identity and Access Management: Developer and Vendor Challenges* (NSA and CISA, October 2023), www.cisa.gov/news-events/alerts/2023/10/04/cisa-and-nsa-release-new-guidance-identity-and-access-management.

[612] Ken Huang, "Agentic AI Identity Management Approach," Cloud Security Alliance, March 11, 2025, https://cloudsecurityalliance.org/blog/2025/03/11/agentic-ai-identity-management-approach.

[613] Matthew K. Carter, *Techniques to Approach Least Privilege* (IDPro, 2022), https://doi.org/10.55621/idpro.88.

[614] "Guide to the Types of Access Control Models," NordLayer, accessed September 2025, https://nordlayer.com/learn/access-control/types-of-access-control/.

[615] Nastaran Farhadighalati, Luis A. Estrada-Jimenez, Sanaz Nikghadam-Hojjati, and Jose Barata, "A Systematic Review of Access Control Models: Background, Existing Research, and Challenges," *IEEE Access* 13 (2025): 17777–17806, https://doi.org/10.1109/ACCESS.2025.3533145.

[616] "Access Control (AC)," CMS CyberGeek, accessed September 2025, https://security.cms.gov/policy-guidance/access-control-ac.

[617] Ken Huang, Vineeth Sai Narajala, John Yeoh et al., "A Novel Zero-Trust Identity Framework for Agentic AI: Decentralized Authentication and Fine-Grained Access Control," arXiv preprint arXiv:2505.19301 (2025), https://doi.org/10.48550/arXiv.2505.19301.

[618] *The Journey to Zero Trust: Microsegmentation in Zero Trust Part One: Introduction and Planning* (CISA, July 29, 2025), www.cisa.gov/resources-tools/resources/microsegmentation-zero-trust-part-one-introduction-and-planning.

[619] Karen Scarfone and Paul Hoffman, *Guidelines on Firewalls and Firewall Policy* (NIST, September 2009), https://doi.org/10.6028/NIST.SP.800-41r1.

[620] "Fundamentals of Cross Domain Solutions," Australian Signals Directorate, December 4, 2019, www.cyber.gov.au/business-government/secure-design/secure-by-design/cross-domain-solutions/fundamentals-of-cross-domain-solutions.

[621] US. Department of Justice, "Justice Department Implements Critical National Security Program to Protect Americans' Sensitive Data from Foreign Adversaries," news release, April 11, 2025, www.justice.gov/opa/pr/justice-department-implements-critical-national-security-program-protect-americans-sensitive.

[622] "International Data Transfers," Data Protection Guide for Small Business, European Data Protection Board, accessed September 2025, www.edpb.europa.eu/sme-data-protection-guide/international-data-transfers_en.

[623] "A Guide to International Transfers," UK Information Commissioner's Office, May 29, 2025, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/international-transfers/international-transfers-a-guide/.

[624] "Data Privacy Framework (DPF) Program Overview," Data Privacy Framework Program, accessed October 12, 2025, www.dataprivacyframework.gov/Program-Overview.

[625] Rogier Creemers and Graham Webster, "Translation: Personal Information Protection Law of the People's Republic of China—Effective Nov. 1, 2021," *DigiChina*, August 20, 2021, revised September 7, 2021, https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/.

[626] Ronaldo Lemos, Natalia Langenegger, Juliana Pacetta Ruiz et al., "Brazilian General Data Protection Law (LGPD, English Translation)," IAPP, last modified October 2020, https://prod.iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/.

[627] Canadian Centre for Cyber Security, "Data Transfer and Upload Protection—ITSAP.40.212," Government of Canada, December 2022, www.cyber.gc.ca/en/guidance/data-transfer-upload-protection-itsap40212.

[628] "Guidance on Privacy and the Use of Commercially Available AI Products," Office of the Australian Information Commissioner, Australian Government, last modified January 17, 2025, www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/guidance-on-privacy-and-the-use-of-commercially-available-ai-products#section-what-are-the-key-privacy-risks-when-using-ai/.

[629] Confederation of European Data Protection Organizations AI Working Group, *Generative AI: The Data Protection Implications* (CEDPO, October 16, 2023), https://cedpo.eu/generative-ai-the-data-protection-implications/.

[630] CybSafe, "STUDY: Almost 40% of Workers Share Sensitive Information with AI Tools, Without Employer's Knowledge," news release, September 26, 2024, www.cybsafe.com/press-releases/study-almost-40-of-workers-share-sensitive-information-with-ai-tools-without-employers-knowledge/.

[631] *From Payrolls to Patents: The Spectrum of Data Leaked into GenAI* (Harmonic, 2024), https://www.harmonic.security/resources/from-payrolls-to-patents-the-spectrum-of-data-leaked-into-genai.

[632] Nick van der Meulen and Barbara H. Wixom, "Bring Your Own AI: How to Balance Risks and Innovation," *MIT Sloan Management Review*, October 3, 2024, https://sloanreview.mit.edu/article/bring-your-own-ai-how-to-balance-risks-and-innovation/.

[633] Osonde A. Osoba, Benjamin Boudreaux, Jessica Saunders, J. Luke Irwin, Pam A. Mueller, and Samantha Cherney, *Algorithmic Equity: A Framework for Social Applications* (RAND, July 11, 2019), www.rand.org/pubs/research_reports/RR2708.html.

[634] Rifat Ara Shams, Didar Zowghi, and Muneera Bano, "AI and the Quest for Diversity and Inclusion: A Systematic Literature Review," *AI and Ethics* 5 (2025): 411–438, https://doi.org/10.1007/s43681-023-00362-w.

[635] Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian, "Manipulating and Mitigating Generative Model Biases Without Retraining," in *Computer Vision – ECCV 2024 Workshops*, eds. A. Del Bue, C. Canton, J. Pont-Tuset, and T. Tommasi (Springer, 2025), https://doi.org/10.1007/978-3-031-92089-9_5.

[636] Rubén González-Sendino, Emilio Serrano, and Javier Bajo, "Mitigating Bias in Artificial Intelligence: Fair Data Generation via Causal Models for Transparent and Explainable Decision-Making," *Future Generation Computer Systems* 155 (2024): 384–401, https://doi.org/10.1016/j.future.2024.02.023.

[637] Zhixuan Chu, Yan Wang, Longfei Li, Zhibo Wang, Zhan Qin, and Kui Ren, "A Causal Explainable Guardrails for Large Language Models," in *CCS '24: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (Association for Computing Machinery, 2024), 1136–1150, https://doi.org/10.1145/3658644.3690217.

[638] "Input Validation Cheat Sheet," OWASP Cheat Sheet Series, accessed September 2025, https://cheatsheetseries.owasp.org/cheatsheets/Input_Validation_Cheat_Sheet.html.

[639] Xiaoyu Zhang, Cen Zhang, Tianlin Li et al., "JailGuard: A Universal Detection Framework for Prompt-based Attacks on LLM Systems," *ACM Transactions on Software Engineering Methodology* (March 2025), https://doi.org/10.1145/3724393.

[640] Amrita Roy Chowdhury, David Glukhov, Divyam Anshumaan et al., "Prεεmpt: Sanitizing Sensitive Prompts for LLMs," arXiv preprint arXiv:2504.05147 (2025), https://doi.org/10.48550/arXiv.2504.05147.

[641] "Injection Prevention Cheat Sheet," OWASP Cheat Sheet Series, accessed September 2025, https://cheatsheetseries.owasp.org/cheatsheets/Injection_Prevention_Cheat_Sheet.html.

[642] Yupei Liu, Yuqi Jia, Runpeng Gengm, Jinyuan Jia, and Neil Zhenqiang Gong, "Formalizing and Benchmarking Prompt Injection Attacks and Defenses," in *33rd USENIX Security Symposium* (USENIX Association, 2025), www.usenix.org/conference/usenixsecurity24/presentation/liu-yupei.

[643] "LLM01:2025 Prompt Injection," GenAI Security Project, OWASP, accessed September 2025, https://genai.owasp.org/llmrisk/llm01-prompt-injection/.

[644] "LLM Prompt Injection Prevention Cheat Sheet," OWASP Cheat Sheet Series, accessed TK, https://cheatsheetseries.owasp.org/cheatsheets/LLM_Prompt_Injection_Prevention_Cheat_Sheet.html.

[645] Jessica Ji, Josh A. Goldstein, and Andrew Lohn, *Controlling Large Language Model Outputs: A Primer* (CSET, December 2023), https://cset.georgetown.edu/publication/controlling-large-language-models-a-primer/.

[646] "LLM09:2025 Misinformation," OWASP, accessed September 2025, https://genai.owasp.org/llmrisk/llm092025-misinformation/.

[647] "LLM05:2025 Improper Output Handling," OWASP, accessed September 2025, https://genai.owasp.org/llmrisk/llm052025-improper-output-handling/.

[648] Yi Dong, Ronghui Mu, Gaojie Jin et al., "Position: Building Guardrails for Large Language Models Requires Systematic Design," *Proceedings of Machine Learning Research* 235 (2024): 11375–11394, https://proceedings.mlr.press/v235/dong24c.html.

[649] Syed Akheel, "Guardrails for Large Language Models: A Review of Techniques and Challenges," *Journal of Artificial Intelligence, Machine Learning and Data Science* 3 (2025): 2504–2512, https://doi.org/10.51219/JAIMLD/syed-arham-akheel/536.

[650] Chen Yueh-Han, Nitish Joshi, Yulin Chen, He He, and Rico Angell, "Monitoring LLM Agents for Sequentially Contextual Harm," OpenReview (March 2025), https://openreview.net/forum?id=LC0XQ6ufbr.

[651] "PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action," Partnership on AI, accessed September 2025, https://syntheticmedia.partnershiponai.org/.

[652] *Contextualizing Deepfake Threats to Organizations* (NSA, FBI, and CISA, 2023), www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3523329/nsa-us-federal-agencies-advise-on-deepfake-threats/.

[653] Siddarth Srinivasan, "Detecting AI Fingerprints: A Guide to Watermarking and Beyond," Brookings, January 4, 2024 www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/.

[654] "Content Credentials: Strengthening Multimedia Integrity in the Generative AI Era," Australian Signals Directorate, January 30, 2025, www.cyber.gov.au/business-government/secure-design/artificial-intelligence/content-credentials-strengthening-multimedia-integrity-in-generative-ai-era.

[655] Shivani Kapania, Stephanie Ballard, Alex Kessler, and Jennifer Wortman Vaughan, "Examining the Expanding Role of Synthetic Data Throughout the AI Development Pipeline," in *FAccT '25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2025), 45–60, https://doi.org/10.1145/3715275.3732005.

[656] Cedric Deslandes Whitney and Justin Norman, "Real Risks of Fake Data: Synthetic Data, Diversity-Washing and Consent Circumvention," in *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2024), 1733–1744, https://doi.org/10.1145/3630106.3659002.

[657] Yonadav Shavit, Sandhini Agarwal, Miles Brundage et al., *Practices for Governing Agentic AI Systems* (OpenAI, December 14, 2023), https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

[658] Javier Garcia and Fernando Fernández, "Safe Exploration of State and Action Spaces in Reinforcement Learning," *Journal of Artificial Intelligence Research* 45 (2012): 515–564, https://doi.org/10.1613/jair.3761.

[659] Alan Chan, Rebecca Salganik, Alva Markelius et al., "Harms from Increasingly Agentic Algorithmic Systems," in *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2023), 651–666, https://doi.org/10.1145/3593013.3594033.

[660] Dan Hendrycks, "Single-Agent Safety," in *Introduction to AI Safety, Ethics, and Society* (Center for AI Safety, 2023), www.aisafetybook.com/.

[661] Zehang Deng, Yongjian Guo, Changzhou Han et al., "AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways," *ACM Computing Surveys* 57, no. 7 (July 2025), https://doi.org/10.1145/3716628.

[662] "LLM06:2025 Excessive Agency," OWASP, accessed September 2025, https://genai.owasp.org/llmrisk/llm062025-excessive-agency/.

[663] Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du et al., "Evaluating Language-Model Agents on Realistic Autonomous Tasks," arXiv preprint arXiv:2312.11671 (2024), https://doi.org/10.48550/arXiv.2312.11671.

[664] Lakshmi Varanasi, "Don't Get Too Excited About AI Agents Yet. They Make a Lot of Mistakes," *Business Insider*, April 17, 2025, www.businessinsider.com/ai-agents-errors-hallucinations-compound-risk-2025-4.

[665] Helen Toner, John Bansemer, Kyle Crichton et al., *Through the Chat Window and into the Real World: Preparing for AI Agents* (CSET, October 2024), https://doi.org/10.51593/20240034.

[666] Jonas C. Ditz, Veronika Lazar, Elmar Lichtmeß et al., "Secure Human Oversight of AI: Exploring the Attack Surface of Human Oversight," arXiv preprint arXiv:2509.12290 (2025), https://doi.org/10.48550/arXiv.2509.12290.

[667] Tobin South, Samuele Marro, Thomas Hardjono et al., "Authenticated Delegation and Authorized AI Agents," arXiv preprint arXiv:2501.09674 (2025), https://doi.org/10.48550/arXiv.2501.09674.

[668] Mary Lynn Garcia, *The Design and Evaluation of Physical Protection Systems*, 2nd ed. (Elsevier, 2007), https://books.google.com/books?hl=en&lr=&id=NDMVuN_4VfIC&oi=fnd&pg=PP1&dq=guide+to+IT+physical+security&ots=I5-ywclWoC&sig=ZGwen7ECZyKPuprXpxeOu1pX-mg.

[669] UK National Protective Security Authority and National Cyber Security Centre, "Data Center Security: Guidance for Owners and Users," National Protective Security Authority, last modified March 17, 2022, www.npsa.gov.uk/system-information-security/data-centre-security/.

[670] "Safety and Functional Safety: IEC 61508 Series," International Electrotechnical Commission, accessed July 2025, www.iec.ch/functional-safety.

[671] *Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data* (OECD Publishing, February 2025), www.oecd.org/en/publications/intellectual-property-issues-in-artificial-intelligence-trained-on-scraped-data_d5241a23-en.html.

[672] *Copyright and Artificial Intelligence, Part 3: Generative AI Training*, prepublication version (U.S. Copyright Office, May 2025), www.copyright.gov/ai/.

[673] Matt Blaszczyk, Geoffrey McGovern, and Karlyn D. Stanley, "Artificial Intelligence Impacts on Copyright Law," RAND, November 20, 2024, www.rand.org/pubs/perspectives/PEA3243-1.html.

[674] Chloe Veltman, "Anthropic Settles with Authors in First-of-its-Kind AI Copyright Infringement Lawsuit," NPR, September 5, 2025, www.npr.org/2025/09/05/nx-s1-5529404/anthropic-settlement-authors-copyright-ai.

[675] *Copyright and Artificial Intelligence, Part 1: Digital Replicas* (U.S. Copyright Office, July 2024), www.copyright.gov/ai/.

[676] An Act to Amend the General Obligations Law, in Relation to Contracts for the Creation and Use of Digital Replicas, S.B. S7676B, New York State Senate (2023–2024), www.nysenate.gov/legislation/bills/2023/S7676/amendment/B.

[677] Contracts Against Public Policy: Personal or Professional Services: Digital Replicas, A.B. 2602, California State Assembly (2023–2024), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2602.

[678] Ensuring Likeness, Voice, and Image Security Act of 2024, Public Chapter No. 588, H.B. 2091, Tennessee General Assembly (2024), https://wapp.capitol.tn.gov/apps/BillInfo/default.aspx?BillNumber=HB2091&GA=113.

[679] Deepak Somaya and Lav R. Varshney. “Embodiment, Anthropomorphism, and Intellectual Property Rights for AI Creations,” in *AIES ’18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery, 2018), 278–283, https://doi.org/10.1145/3278721.3278754.

[680] Zhengyuan Jiang, Moyang Guo, Yuepeng Hu, Yupu Wang, and Neil Zhenqiang Gong, “Watermark-Based Attribution of AI-Generated Content,” arXiv preprint arXiv:2404.04254 (2024), https://doi.org/10.48550/arXiv.2404.04254.

[681] *Copyright and Artificial Intelligence, Part 2: Copyrightability* (U.S. Copyright Office, January 2025), www.copyright.gov/ai/.

[682] U.S. Patent and Trademark Office, “Inventorship Guidance for AI-Assisted Inventions,” 89 FR 10043 (February 13, 2024), www.federalregister.gov/documents/2024/02/13/2024-02623/inventorship-guidance-for-ai-assisted-inventions.

[683] Kevin J. Hickey and Christopher T. Zirpoli, "Artificial Intelligence and Patent Law,” Congressional Research Service, Library of Congress, December 12, 2024, www.congress.gov/crs-product/LSB11251.

[684] “Worker Rights,” U.S. Department of Labor, accessed September 2025, www.dol.gov/agencies/whd/workers.

[685] *Artificial Intelligence and Worker Well-Being: Principles and Best Practices for Developers and Employers* (U.S. Department of Labor, October 2024), www.privacysecurityacademy.com/wp-content/uploads/2024/10/US-Dept.-of-Labor-AI-Principles-Best-Practices.pdf.

[686] Sonam Jindal, “AI and Human Rights: Protecting Data Workers,” Partnership on AI, May 1, 2025, https://partnershiponai.org/ai-and-human-rights-protecting-data-workers/.

[687] “Humans in the AI Loop: The Data Labelers Behind Some of the Most Powerful LLMs’ Training Datasets,” Privacy International, August 15, 2024, https://privacyinternational.org/explainer/5357/humans-ai-loop-data-labelers-behind-some-most-powerful-llms-training-datasets.

[688] International Labour Organization, International Social Security Association, and OECD, *Providing Adequate and Sustainable Social Protection for Workers in the Gig and Platform Economy* (G20, January 2023), https://g20ewgportal.org/documents/providing-adequate-and-sustainable-social-protection-for-workers-in-the-gig-and-platform-economy.

[689] Moon Hwan Lee, “Reimagining Workers’ Rights in the Gig Economy: Bridging the Gap Between Independent Contractors and Employees,” New York State Bar Association, August 5, 2025, https://nysba.org/reimagining-workers-rights-in-the-gig-economy-bridging-the-gap-between-independent-contractors-and-employees/.

[690] Sachin R. Pendse, Darren Gergle, Rachel Kornfield et al., “When Testing AI Tests Us: Safeguarding Mental Health on the Digital Frontlines,” in *FAccT ’25: Proceedings of the 2025 ACM*

Conference on Fairness, Accountability, and Transparency (Association for Computing Machinery, 2025), 1793–1804, https://doi.org/10.1145/3715275.3732120.

[691] Brandon Dang, Martin J. Riedl, and Matthew Lease, "But Who Protects the Moderators? The Case of Crowdsourced Image Moderation," arXiv preprint arXiv:1804.10999 (2018), https://doi.org/10.48550/arXiv.1804.10999.

[692] Keith A. Markel, Alana R. Mildner, and Jessica L. Lipson, "AI and Employee Privacy: Important Considerations for Employers," Reuters, September 29, 2023, www.reuters.com/legal/legalindustry/ai-employee-privacy-important-considerations-employers-2023-09-29/.

[693] Consumer Financial Protection Bureau, "CFPB Takes Action to Curb Unchecked Worker Surveillance," news release, October 24, 2024, www.consumerfinance.gov/about-us/newsroom/cfpb-takes-action-to-curb-unchecked-worker-surveillance/.

[694] Ezra Awumey, Sauvik Das, and Jodi Forlizzi, "A Systematic Review of Biometric Monitoring in the Workplace: Analyzing Socio-technical Harms in Development, Deployment and Use," in *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2024), 920–932, https://doi.org/10.1145/3630106.3658945.

[695] Kat Roemmich, Florian Schaub, and Nazanin Andalibi, "Emotion AI at Work: Implications for Workplace Surveillance, Emotional Labor, and Emotional Privacy," in *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2023), https://doi.org/10.1145/3544548.3580950.

[696] Anna Lena Hunkenschroer and Alexander Kriebitz, "Is AI Recruiting (Un)Ethical? A Human Rights Perspective on the Use of AI for Hiring," *AI and Ethics* 3 (2023): 199–213, https://doi.org/10.1007/s43681-022-00166-4.

[697] "Background Dossiers and Algorithmic Scores for Hiring, Promotion, and Other Employment Decisions," CFPB, protection circular 2024-06, October 24, 2024, www.consumerfinance.gov/compliance/circulars/consumer-financial-protection-circular-2024-06-background-dossiers-and-algorithmic-scores-for-hiring-promotion-and-other-employment-decisions/.

[698] *Network Infrastructure Security Guide* (NSA, October 2023), www.nsa.gov/Press-Room/News-Highlights/Article/Article/2949885/nsa-details-network-infrastructure-best-practices/.

[699] Ramaswamy Chandramouli, *Guide to a Secure Enterprise Network Landscape* (NIST, November 2022), https://doi.org/10.6028/NIST.SP.800-215.

[700] Sasidhar Chennamsetty and Sai Ankith Averineni, "From Network Segmentation to AI-Driven Zero Trust: A Systematic Survey of Micro-segmentation Technologies," in *2025 7th International Conference on Electronics and Communication, Network and Computer Technology* (IEEE, 2025), 111–117, https://doi.org/10.1109/ECNCT66493.2025.11172491.

[701] Karen Scarfone and Peter Mell, *Guide to Intrusion Detection and Prevention Systems (IDPS)* (NIST, February 2007), https://doi.org/10.6028/NIST.SP.800-94.

[702] Neha Gupta, Vinita Jindal, and Punam Bedi, "A Survey on Intrusion Detection and Prevention Systems," *SN Computer Science* 4, no. 439 (2023), https://doi.org/10.1007/s42979-023-01926-7.

[703] Amir Javadpour, Forough Ja'fari, Tarik Taleb, Mohammad Shojafar, and Chafika Benzaïd, "A Comprehensive Survey on Cyber Deception Techniques to Improve Honeypot Performance," *Computers & Security* 140 (2024), https://doi.org/10.1016/j.cose.2024.103792.

[704] Canadian Centre for Cyber Security, "Security Considerations for Edge Devices," Government of Canada, February 2025, www.cyber.gc.ca/en/guidance/security-considerations-edge-devices-itsm80101.

[705] "Device Security Guidance," UK National Cyber Security Centre, last modified May 13, 2025, www.ncsc.gov.uk/collection/device-security-guidance.

[706] Gema Howell, Joshua M. Franklin, Vincent Sritapan, Murugiah Souppaya, and Karen Scarfone, *Guidelines for Managing the Security of Mobile Devices in the Enterprise* (NIST, May 2023), https://doi.org/10.6028/NIST.SP.800-124r2.

[707] Karen Scarfone, Wayne Jansen, and Miles Tracy, *Guide to General Server Security* (NIST, July 2008), https://doi.org/10.6028/NIST.SP.800-123.

[708] Canadian Centre for Cyber Security, "Protect Your Organization from Malware (ITAP.00.057)," Government of Canada, July 2022, www.cyber.gc.ca/en/guidance/protect-your-organization-malware-itsap00057/.

[709] Canadian Centre for Cyber Security, "Preventative Security Tools (ITSAP:00:0058)," Government of Canada, May 2024, www.cyber.gc.ca/en/guidance/preventative-security-tools-itsap00058.

[710] "No-Cost Cybersecurity Services and Tools," CISA, accessed September 2025, www.cisa.gov/resources-tools/resources/free-cybersecurity-services-and-tools.

[711] Rohani Rohan, Borworn Papasratorn, Wichian Chutimaskul, Jari Hautamäki, Suree Funilkul, and Debajyoti Pal, "Enhancing Cybersecurity Resilience: A Comprehensive Analysis of Human Factors and Security Practices Aligned with the NIST Cybersecurity Framework," in *IAIT '23: Proceedings of the 13th International Conference on Advances in Information Technology* (Association for Computing Machinery, 2023), https://doi.org/10.1145/3628454.3629472.

[712] "Module 1: Base Cybersecurity for Personal Computers and Mobile Devices," CISA, accessed September 2025, www.cisa.gov/audiences/high-risk-communities/projectupskill/module1.

[713] "A09:2021—Security Logging and Monitoring Failures," OWASP Top 10, 2021, https://owasp.org/Top10/A09_2021-Security_Logging_and_Monitoring_Failures/.

[714] Bernie Lantz, Rob Hall, and Jason Couraud, "Locking Down Log Files: Enhancing Network Security by Protecting Log Files," *Issues in Information Systems* 7, no. 2 (2006), https://doi.org/10.48009/2_iis_2006_43-47.

[715] Canadian Centre for Cyber Security, "Tips for Backing Up Your Information (ITSAP.40.002)," Government of Canada, June 2024, www.cyber.gc.ca/en/guidance/tips-backing-your-information-itsap40002.

[716] "Data Backup," MITRE, last modified December 10, 2024, https://attack.mitre.org/mitigations/M1053/.

[717] Paul Ruggiero and Matthew A. Heckathorn, *Data Backup Options* (U.S. Computer Emergency Readiness Team, 2012), www.cisa.gov/sites/default/files/publications/data_backup_options.pdf.

[718] National Cybersecurity Center of Excellence, *Protecting Data from Ransomware and Other Data Loss Events* (NIST, April 2020), https://csrc.nist.gov/pubs/other/2020/04/24/protecting-data-from-ransomware-and-other-data-los/final.

[719] William Fisher, R. Eugene Craft, Michael Ekstrom, Julian Sexton, and John Sweetnam, *Data Confidentiality: Identifying and Protecting Assets Against Data Breaches* (NIST, February 2024), https://doi.org/10.6028/NIST.SP.1800-28.

[720] Erika McCallister, Tim Grance, and Karen Scarfone, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)* (NIST, April 2010), https://doi.org/10.6028/NIST.SP.800-122.

[721] "Quantum-Readiness: Migration to Post-quantum Cryptography," CISA, August 21, 2023, www.cisa.gov/resources-tools/resources/quantum-readiness-migration-post-quantum-cryptography.

[722] Information Technology Laboratory, "Module-Lattice-Based Key-Encapsulation Mechanism Standard," NIST, August 13, 2024, https://doi.org/10.6028/NIST.FIPS.203.

[723] "LLM02:2025 Sensitive Information Disclosure," OWASP, accessed September 2025, https://genai.owasp.org/llmrisk/llm022025-sensitive-information-disclosure/.

[724] Li Hu, Anli Yan, Hongyang Yan et al. "Defenses to Membership Inference Attacks: A Survey," *ACM Computing Surveys* 56, no. 4 (April 2024), https://doi.org/10.1145/3620667.

[725] "ML04:2023 Membership Inference Attack," OWASP, accessed September 2025, https://owasp.org/www-project-machine-learning-security-top-10/docs/ML04_2023-Membership_Inference_Attack.html.

[726] Zhanke Zhou, Jianing Zhu, Fengfei Yu et al., "Model Inversion Attacks: A Survey of Approaches and Countermeasures," arXiv preprint arXiv:2411.10023 (2024), https://doi.org/10.48550/arXiv.2411.10023.

[727] "ML03:2023 Model Inversion Attack," OWASP, accessed September 2025, https://owasp.org/www-project-machine-learning-security-top-10/docs/ML03_2023-Model_Inversion_Attack.html.

[728] Information Technology Laboratory, “Stateless Hash-Based Digital Signature Standard,” NIST, August 13, 2024, https://doi.org/10.6028/NIST.FIPS.205.

[729] Information Technology Laboratory, “Module-Lattice-Based Digital Signature Standard,” NIST, August 13, 2024, https://doi.org/10.6028/NIST.FIPS.204.

[730] “Understanding Digital Signatures,” CISA, February 1, 2021, www.cisa.gov/news-events/news/understanding-digital-signatures/.

[731] Gopalan Sivathanu, Charles P. Wright, and Erez Zadok, “Ensuring Data Integrity in Storage: Techniques and Applications,” in *StorageSS ’05: Proceedings of the 2005 ACM Workshop on Storage Security and Survivability* (Association for Computing Machinery, 2005), 26–36, https://doi.org/10.1145/1103780.1103784.

[732] Yihui Dong, Le Sun, Dengzhi Liu, Meng Feng, and Tiantian Miao, “A Survey on Data Integrity Checking in Cloud," in *2018 1st International Cognitive Cities Conference* (IEEE, 2018), 109–113, https://doi.org/10.1109/IC3.2018.00031.

[733] NSA, CISA, FBI et al., *AI Data Security: Best Practices for Securing Data Used to Train and Operate AI Systems* (CISA, May 2025), www.cisa.gov/resources-tools/resources/ai-data-security-best-practices-securing-data-used-train-operate-ai-systems.

[734] Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis et al., “Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning,” *ACM Computing Surveys* 55, no. 13s (December 2023), https://doi.org/10.1145/3585385.

[735] “ML02:2023 Data Poisoning Attack,” OWASP, accessed September 2025, https://owasp.org/www-project-machine-learning-security-top-10/docs/ML02_2023-Data_Poisoning_Attack.html.

[736] “LLM04:2025 Data and Model Poisoning,” OWASP, accessed September 2025, https://genai.owasp.org/llmrisk/llm042025-data-and-model-poisoning/.

[737] “ML08:2023 Model Skewing,” OWASP, accessed September 2025, https://owasp.org/www-project-machine-learning-security-top-10/docs/ML08_2023-Model_Skewing.html.

[738] Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott, *Security AI Model Weights* (RAND, May 30, 2024), www.rand.org/pubs/research_reports/RRA2849-1.html.

[739] “ML05:2023 Model Theft,” OWASP, accessed September 2025, https://owasp.org/www-project-machine-learning-security-top-10/docs/ML05_2023-Model_Theft.html.

[740] Zhichuang Sun, Ruimin Sun, Long Lu, and Alan Mislove, “Mind Your Weight(s): A Large-Scale Study on Insufficient Machine Learning Model Protection in Mobile Apps,” in *30th USENIX Security*

Symposium (USENIX Association, 2021), www.usenix.org/conference/usenixsecurity21/presentation/sun-zhichuang.

[741] Jingwen Ye, Yining Mao, Jie Song, Xinchao Wang, Cheng Jin, and Mingli Song, "Safe Distillation Box," *Proceedings of the AAAI Conference on Artificial Intelligence* 36, no. 3 (2022), 3117–3124, https://doi.org/10.1609/aaai.v36i3.20219.

[742] Kaixiang Zhao, Lincan Li, Kaize Ding, Neil Zhenqiang Gong, Yue Zhao, and Yushun Dong, "A Survey on Model Extraction Attacks and Defenses for Large Language Models," in *KDD '25: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2025), 6227–6236, https://doi.org/10.1145/3711896.3736573.

[743] Hervé Chabanne, Jean-Luc Danger, Linda Guiga, and Ulrich Kühne, "Side Channel Attacks for Architecture Extraction of Neural Networks," *CAAI Transactions on Intelligence Technology* 6, no. 1 (2021): 3–16, https://doi.org/10.1049/cit2.12026.

[744] Younghan Lee, Sohee Jun, Yungi Cho, Woorim Han, Hyungon Moon, and Yunheung Paek, "Precise Extraction of Deep Learning Models via Side-Channel Attacks on Edge/Endpoint Devices," in *Computer Security – ESORICS 2022* (Springer, 2022), https://doi.org/10.1007/978-3-031-17143-7_18.

[745] "ML10:2023 Model Poisoning," OWASP, accessed September 2025, https://owasp.org/www-project-machine-learning-security-top-10/docs/ML10_2023-Model_Poisoning.html.

[746] "ML09:2023 Output Integrity Attack," OWASP, accessed September 2025, https://owasp.org/www-project-machine-learning-security-top-10/docs/ML09_2023-Output_Integrity_Attack.html.

[747] Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Anderson, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* (NIST, January 2024), https://doi.org/10.6028/NIST.AI.100-2e2023.

[748] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay, "A Survey on Adversarial Attacks and Defences," *CAAI Transactions on Intelligent Technology* 6, no. 1 (March 2021): 25–45, https://doi.org/10.1049/cit2.12028.

[749] Shuai Zhou, Chi Liu, Dayong Ye, Tianqing Zhu, Wanlei Zhou, and Philip S. Yu, "Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity," *ACM Computing Surveys* 55, no. 8 (August 2023), https://doi.org/10.1145/3547330.

[750] Shunyao Wang, Ryan K. L. Ko, Guangdong Bai, Naipeng Dong, Taejun Choi, and Yanjun Zhang, "Evasion Attack and Defense on Machine Learning Models in Cyber-Physical Systems: A Survey," *IEEE Communications Surveys & Tutorials* 26, no. 2 (2024): 930–966, https://doi.org/10.1109/COMST.2023.3344808.

[751] Raja Muthalagu, Jasmita Malik, Pranav M. Pawar, "Detection and Prevention of Evasion Attacks on Machine Learning Models," *Expert Systems with Applications* 266 (2025), https://doi.org/10.1016/j.eswa.2024.126044.

[752] "ML01:2023 Input Manipulation Attack," OWASP, accessed September 2025, https://owasp.org/www-project-machine-learning-security-top-10/docs/ML01_2023-Input_Manipulation_Attack.html.

[753] Salman Rahman, Liwei Jiang, James Shiffer et al., "X-Teaming: Multi-turn Jailbreaks and Defenses with Adaptive Multi-agents," arXiv preprint arXiv:2504.13203 (2025), https://doi.org/10.48550/arXiv.2504.13203.

[754] "LLM08:2025 Vector and Embedding Weaknesses," OWASP, accessed September 2025, https://genai.owasp.org/llmrisk/llm082025-vector-and-embedding-weaknesses/.

[755] Congzheng Song and Ananth Raghunathan, "Information Leakage in Embedding Models," in *CCS '20: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security* (Association for Computing Machinery, 2020), 377–390, https://doi.org/10.1145/3372297.3417270.

[756] Antonios Tragoudaras, Theofanis Aslanidis, Emmanouil Georgios Lionis, Marina Orozco González, and Panagiotis Eustratiadis, "Information Leakage of Sentence Embeddings via Generative Embedding Inversion Attacks," in *SIGIR '25: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Association for Computing Machinery, 2025), 3234–3243, https://doi.org/10.1145/3726302.3730303.

[757] Linke Song, Zixuan Pang, Wenhao Wang et al., "The Early Bird Catches the Leak: Unveiling Timing Side Channels in LLM Serving Systems," arXiv preprint arXiv:2409.20002 (2024), https://doi.org/10.48550/arXiv.2409.20002.

[758] Andrew Adiletta and Berk Sunar, "Spill the Beans: Exploiting CPU Cache Side-Channels to Leak Tokens from Large Language Models," arXiv preprint arXiv:2505.00817 (2025), https://doi.org/10.48550/arXiv.2505.00817.

[759] Zibo Gao, Junjie Hu, Feng Guo et al., "I Know What You Said: Unveiling Hardware Cache Side-Channels in Local Large Language Model Inference," arXiv preprint arXiv:2505.06738 (2025), https://doi.org/10.48550/arXiv.2505.06738.

[760] Marcin Chrapek, Marcin Copik, Etienne Mettaz, and Torsten Hoefler, "Confidential LLM Inference: Performance and Cost Across CPU and GPU TEEs," arXiv preprint arXiv:2509.18886 (2025), https://doi.org/10.48550/arXiv.2509.18886.

[761] Arunava Chaudhuri, Shubhi Shukla, Sarani Bhattacharya, and Debdeep Mukhopadhyay, "Secured and Privacy-Preserving GPU-Based Machine Learning Inference in Trusted Execution Environment: A Comprehensive Survey," in *2025 17th International Conference on COMmunication Systems and NETworks* (IEEE, 2025), 207–216, https://doi.org/10.1109/COMSNETS63942.2025.10885734.

[762] Stavros Volos, Kapil Vaswani, and Rodrigo Bruno, "Graviton: Trusted Execution Environments on GPUs," in *13th USENIX Symposium on Operating Systems Design and Implementation* (USENIX Association, 2018), 681–696, www.usenix.org/conference/osdi18/presentation/volos.

[763] Gobikrishna Dhanuskodi, Sudeshna Guha, Vidhya Krishnan et al., "Creating the First Confidential GPUs," *Communications of the ACM* 67 (2024), https://dl.acm.org/doi/pdf/10.1145/3626827.

[764] Shulin Fan, Zhichao Hua, Yubin Xia, and Haibo Chen, "XpuTEE: A High-Performance and Practical Heterogeneous Trusted Execution Environment for GPUs," *ACM Transactions on Computer Systems* 43, no. 1–2 (May 2025), https://doi.org/10.1145/3719653.

[765] Aritra Dhar, Clément Thorens, Lara Magdalena Lazier, and Lukas Cavigelli, "GuardAIn: Protecting Emerging Generative AI Workloads on Heterogeneous NPU," in *2025 IEEE Symposium on Security and Privacy* (IEEE, 2025), 4155–4172, https://doi.org/10.1109/SP61157.2025.00221.

[766] Maria I. Mera Collantes, Zahra Ghodsi, and Siddharth Garg, "SafeTPU: A Verifiably Secure Hardware Accelerator for Deep Neural Networks," in *2020 IEEE 38th VLSI Test Symposium* (IEEE, 2020), 1–6, https://doi.org/10.1109/VTS48691.2020.9107564.

[767] Reshma Lal, James B. Anderson, and Andrew Jackson, "Data Processing Unit's Entry into Confidential Computing," in *HASP '23: Proceedings of the 12th International Workshop on Hardware and Architectural Support for Security and Privacy* (Association for Computing Machinery, 2023), 56–63, https://doi.org/10.1145/3623652.3623670.

[768] Bochuan Cao, Changjiang Li, Yuanpu Cao, Yameng Ge, Ting Wang, and Jinghui Chen, "You Can't Steal Nothing: Mitigating Prompt Leakages in LLMs via System Vectors," arXiv preprint arXiv:2509.21884 (2025), https://doi.org/10.48550/arXiv.2509.21884.

[769] Divyansh Agarwal, Alexander Fabbri, Ben Risher, Philippe Laban, Shafiq Joty, and Chien-Sheng Wu, "Prompt Leakage Effect and Mitigation Strategies for Multi-turn LLM Applications," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2024), 1255–1275, https://doi.org/10.18653/v1/2024.emnlp-industry.94.

[770] "LLM07:2025 System Prompt Leakage," OWASP, accessed September 2025, https://genai.owasp.org/llmrisk/llm072025-system-prompt-leakage/.

[771] "CWE-400: Uncontrolled Resource Consumption," Common Weakness Enumeration, MITRE, last modified January 21, 2026, https://cwe.mitre.org/data/definitions/400.html.

[772] Theyazn H. H. Aldhyani and Hasan Alkahtani, "Artificial Intelligence Algorithm-Based Economic Denial of Sustainability Attack Detection Systems: Cloud Computing Environments," *Sensors* 22, no. 13 (2022): 4685, https://doi.org/10.3390/s22134685.

[773] Conor Bronsdon, "How Unbounded Consumption Attacks on LLMs Cost Companies Millions," Galileo.ai, June 27, 2025, https://galileo.ai/blog/prevent-llm-unbounded-consumption/.

[774] Yuanhe Zhang, Xinyue Wang, Haoran Gao et al., "PD3F: A Pluggable and Dynamic DoS-Defense Framework Against Resource Consumption Attacks Targeting Large Language Models," arXiv preprint arXiv:2505.18680 (2025), https://arxiv.org/abs/2505.18680.

[775] "LLM10:2025 Unbounded Consumption," OWASP, accessed September 2025, https://genai.owasp.org/llmrisk/llm102025-unbounded-consumption/.

[776] "What Is Capacity Planning?," IBM, accessed August 2025, www.ibm.com/think/topics/capacity-planning.

[777] "Architecture Strategies for Capacity Planning," Microsoft, last modified August 7, 2025, https://learn.microsoft.com/en-us/azure/well-architected/performance-efficiency/capacity-planning.

[778] Mohammad Masdari and Afsane Khoshnevis, "A Survey and Classification of the Workload Forecasting Methods in Cloud Computing," *Cluster Computing* 23 (2020): 2399–2424, https://doi.org/10.1007/s10586-019-03010-3.

[779] Seyedehmehrnaz Mireslami, Logan Rakai, Mea Wang, and Behrouz Homayoun Far, "Dynamic Cloud Resource Allocation Considering Demand Uncertainty," *IEEE Transactions on Cloud Computing* 9, no. 3 (2021): 981–994, https://doi.org/10.1109/TCC.2019.2897304.

[780] Canadian Centre for Cyber Security, "Defending Against Distributed Denial of Service (DDoS) Attacks—ITSM.80.110," Government of Canada, February 2024, www.cyber.gc.ca/en/guidance/defending-against-distributed-denial-service-ddos-attacks-itsm80110.

[781] "Secure Development and Deployment Guidance: Protect Your Code Repository," UK National Cyber Security Centre, February 20, 2019, www.ncsc.gov.uk/collection/developers-collection/principles/protect-your-code-repository/.

[782] Alexander Krause, Jan H. Klemmer, Nicolas Huaman, Dominik Wermke, Yasemin Acar, and Sascha Fahl, "Pushed by Accident: A Mixed-Methods Study on Strategies of Handling Secret Information in Source Code Repositories," in *32nd USENIX Security Symposium* (USENIX Association, 2023), www.usenix.org/conference/usenixsecurity23/presentation/krause.

[783] "Ongoing Targeting of Online Code Repositories," Australian Signals Directorate, September 19, 2025, www.cyber.gov.au/about-us/view-all-content/alerts-and-advisories/ongoing-targeting-of-online-code-repositories.

[784] David De Cremer and Garry Kasparov, "AI Should Augment Human Intelligence, Not Replace It," *Harvard Business Review*, March 18, 2021, https://hbr.org/2021/03/ai-should-augment-human-intelligence-not-replace-it.

[785] Rohan Alur, Loren Laine, Darrick Li, Manish Raghavan, Devavrat Shah, and Dennis Shung, "Auditing for Human Expertise," *Advances in Neural Information Processing Systems* 36 (2023): 79439–79468, https://proceedings.neurips.cc/paper_files/paper/2023/hash/fb44a668c2d4bc984e9d6ca261262cbb-Abstract-Conference.html.

[786] Isabelle Hau and Rebecca Winthrop, "What Happens When AI Chatbots Replace Real Human Connection," Brookings, July 2, 2025, www.brookings.edu/articles/what-happens-when-ai-chatbots-replace-real-human-connection/.

[787] Cecilia Ka Yuk Chan and Louisa H. Y. Tsi, "Will Generative AI Replace Teachers in Higher Education? A Study of Teacher and Student Perceptions," *Studies in Educational Evaluation* 83 (2024), https://doi.org/10.1016/j.stueduc.2024.101395.

[788] Isabella Loaiza and Roberto Rigobon, "The EPOCH of AI: Human-Machine Complementarities at Work," MIT Sloan, research paper no. 7236-24, November 21, 2024, http://dx.doi.org/10.2139/ssrn.5028371.

[789] Karim Lakhani, "AI Won't Replace Humans—But Humans with AI Will Replace Humans Without AI," *Harvard Business Review*, August 4, 2023, https://hbr.org/2023/08/ai-wont-replace-humans-but-humans-with-ai-will-replace-humans-without-ai.

[790] Ala Yankouskaya, Magnus Liebherr, and Raian Ali, "Can ChatGPT Be Addictive? A Call to Examine the Shift from Support to Dependence in AI Conversational Large Language Models," *Human-Centric Intelligent Systems* 5 (February 17, 2025): 77–89, https://doi.org/10.1007/s44230-025-00090-w.

[791] Artur Klingbeil, Cassandra Grützner, and Philipp Schreck, "Trust and Reliance on AI—An Experimental Study on the Extent and Costs of Overreliance on AI," *Computers in Human Behavior* 160 (2024), https://doi.org/10.1016/j.chb.2024.108352.

[792] Maèva Flayelle, Damien Brevers, Daniel L. King, Pierre Maurage, José C. Perales, and Joël Billieux, "A Taxonomy of Technology Design Features that Promote Potentially Addictive Online Behaviours," *Nature Reviews Psychology* 2 (2023): 136–150, https://doi.org/10.1038/s44159-023-00153-4.

[793] Alicia DeVrio, Myra Cheng, Lisa Egede, Alexandra Olteanu, and Su Lin Blodgett, "A Taxonomy of Linguistic Expressions that Contribute to Anthropomorphism of Language Technologies," in *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2025), https://doi.org/10.1145/3706598.3714038.

[794] Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser, "All Too Human? Mapping and Mitigating the Risks from Anthropomorphic AI," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (2024): 13-26, https://doi.org/10.1609/aies.v7i1.31613.

[795] Takuya Maeda and Anabel Quan-Haase, "When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design," in *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2024), 1068–1077, https://doi.org/10.1145/3630106.3658956.

[796] Lujain Ibrahim, Luc Rocher, Ana Valdivia, "Characterizing and Modeling Harms from Interactions with Design Patterns in AI Interfaces," arXiv preprint arXiv:2404.11370 (2024), https://doi.org/10.48550/arXiv.2404.11370.

[797] Iason Gabriel, Arianna Manzini, Geoff Keeling et al., "The Ethics of Advanced AI Assistants," arXiv preprint arXiv:2404.16244 (2024), https://doi.org/10.48550/arXiv.2404.16244.

[798] Jared Moore, Declan Grabb, William Agnew et al., "Expressing Stigma and Inappropriate Responses Prevents LLMs from Safely Replacing Mental Health Providers," in *FAccT '25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2025), 599–627, https://doi.org/10.1145/3715275.3732039.

[799] Mohit Chandra, Suchismita Naik, Denae Ford et al., "From Lived Experience to Insight: Unpacking the Psychological Risks of Using AI Conversational Agents," in *FAccT '25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2025), 975–1004, https://doi.org/10.1145/3715275.3732063.

[800] Tara Matthews, Elie Bursztein, Patrick Gage Kelley et al., "Supporting the Digital Safety of At-Risk Users: Lessons Learned from 9+ Years of Research and Training," *ACM Transactions on Computer-Human Interaction* 32, no. 3 (June 2025), https://doi.org/10.1145/3716382.

[801] Chunyan Ding, "The Impact of Artificial Intelligence on Vulnerable Individuals: Three Principles," in *Artificial Intelligence and the Future of Human Relations*, eds. Yanto Chandra and Ruiping Fan (Springer, 2025), https://doi.org/10.1007/978-981-96-7185-4_14.

[802] Ge Wang, Jun Zhao, Max Van Kleek, and Nigel Shadbolt, "Informing Age-Appropriate AI: Examining Principles and Practices of AI for Children," in *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2022), https://doi.org/10.1145/3491102.3502057.

[803] *Me, Myself and AI Research: Understanding and Safeguarding Children's Use of AI Chatbots* (InternetMatters.org, July 2025), www.internetmatters.org/hub/research/me-myself-and-ai-chatbot-research/.

[804] *Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions* (Common Sense Media, 2025), www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions.

[805] Dace Civča, Dzintra Atstāja, and Viktor Koval, "Business Continuity Plan Testing Methods in an International Company," in *Knowledge Management Competence for Achieving Competitive Advantage of Professional Growth and Development*, eds. Dzintra Atstāja and Viktor Koval (BA School of Business and Finance, 2021), https://suem.edu.ua/storage/doc/mizhnarodna-spivpracya/latvia-2021.pdf#page=341.

[806] Anna Olson and Jamie Anderson, "Resiliency Scoring for Business Continuity Plans," *Journal of Business Continuity & Emergency Planning* 10, no. 1 (2016): 31–43, www.ingentaconnect.com/content/hsp/jbcep/2016/00000010/00000001/art00004.

[807] Gwen White and Sadie Liptak, "Small Business Continuity and Disaster Recovery Plans Using AI and ChatGPT," *Issues in Information Systems* 25, no. 1 (2021), www.exhibit.xavier.edu/cgi/viewcontent.cgi?article=1265&context=curca.

[808] Gideon N. Angafor, Iryna Yevseyeva, and Ying He, "Game-Based Learning: A Review of Tabletop Exercises for Cybersecurity Incident Response Training," *Security and Privacy* 3, no. 6 (2020), https://doi.org/10.1002/spy2.126.

[809] "CISA Tabletop Exercise Packages," CISA, accessed September 2025, www.cisa.gov/resources-tools/services/cisa-tabletop-exercise-packages/.

[810] Mojisola Aderonke Ojuri, "Measuring Software Resilience: A QA Approach to Cybersecurity Incident Response Readiness," *Multidisciplinary Innovations & Research Analysis* 2, no. 4 (2021), http://openviewjournal.com/index.php/mira/article/view/32.

[811] Anas Amayreh, Mohammad A. Ta'Amnha, Ihab K. Magableh, Maher H. Mahrouq, and Salsabila Aisyah Alfaiza, "Exploring the Impact of AI on Employee Self-Competence Performance Key Variables and Outcomes," *Discover Sustainability* 6 (2025), https://doi.org/10.1007/s43621-025-01438-9.

[812] Maria Arshad, Kamran Hameed, and Hafiz Muhammad Naeem, "Exploring the Role of AI in Shaping Human Competencies and Workforce Development," *Social Science Multidisciplinary Review* 3, no. 1 (2025), https://doi.org/10.69591/ssmr.vol03.no01/003.

[813] André Markus, Astrid Carolus, Carolin Wienrich, "Objective Measurement of AI Literacy: Development and Validation of the AI Competency Objective Scale (AICOS)," *Computers and Education: Artificial Intelligence* 9 (2025), https://doi.org/10.1016/j.caeai.2025.100485.

[814] Joel Dawson and J. Todd McDonald, "Improving Penetration Testing Methodologies for Security-Based Risk Assessment," in *2016 Cybersecurity Symposium* (IEEE, 2016), 51–58, https://doi.org/10.1109/CYBERSEC.2016.016.

[815] Nicholas Athanasiades, Randal Abler, John Levine, H. Owen, and G. Riley, "Intrusion Detection Testing and Benchmarking Methodologies," in *First IEEE International Workshop on Information Assurance* (IEEE, 2003), 63–72, https://doi.org/10.1109/IWIAS.2003.1192459.

[816] Aaron Williams, Alessia Michela Di Campi, Elie Saad et al., "Web Security Testing Guide – Stable," OWASP Foundation, accessed October 2025, https://owasp.org/www-project-web-security-testing-guide/stable/.

[817] Chris Greamo and Anup Ghosh, "Sandboxing and Virtualization: Modern Tools for Combating Malware," *IEEE Security & Privacy* 9, no. 2 (2011): 79–82, https://doi.org/10.1109/MSP.2011.36.

[818] Nesara Dissanayake, Asangi Jayatilaka, Mansooreh Zahedi, and M. Ali Babar, "Software Security Patch Management—A Systematic Literature Review of Challenges, Approaches, Tools and Practice," *Information and Software Technology* 144 (2022), https://doi.org/10.1016/j.infsof.2021.106771.

[819] David Geer, "How to Review and Test Backup Procedures to Ensure Data Restoration," CSO, March 29, 2016, www.csoonline.com/article/555449/how-to-review-and-test-backup-procedures-to-ensure-data-restoration.html.

[820] Jason Thomas and Gordon Galligher, "Improving Backup System Evaluations in Information Security Risk Assessments to Combat Ransomware," *Computer and Information Science* 11, no. 1 (2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3095629.

[821] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence* (NIST, 2022), https://doi.org/10.6028/NIST.SP.1270.

[822] Nengfeng Zhou, Zach Zhang, Vijayan N. Nair, Harsh Singhal, and Jie Chen, "Bias, Fairness and Accountability with Artificial Intelligence and Machine Learning Algorithms," *International Statistical Review* 90, no. 3 (2022): 468–480, https://doi.org/10.1111/insr.12492.

[823] Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel, "The Measure and Mismeasure of Fairness," *Journal of Machine Learning Research* 24 (2023): 1–117, http://jmlr.org/papers/v24/22-1511.html.

[824] Feng Chen, Liqin Wang, Julie Hong, Jiaqi Jiang, and Li Zhou, "Unmasking Bias in Artificial Intelligence: A Systematic Review of Bias Detection and Mitigation Strategies in Electronic Health Record–Based Models," *Journal of the American Medical Informatics Association* 31, no. 5 (2024): 1172–1183, https://doi.org/10.1093/jamia/ocae060.

[825] Anthony M. Barrett, Krystal Jackson, Evan R. Murphy, Nada Madkour, and Jessica Newman, *Benchmark Early and Red Team Often* (UC Berkeley Center for Long-Term Cybersecurity, 2024), https://cltc.berkeley.edu/publication/benchmark-early-and-red-team-often-a-framework-for-assessing-and-managing-dual-use-hazards-of-ai-foundation-models/.

[826] María Victoria Carro, Denise Alejandra Mester, Francisca Gauna Selasco et al., "A Conceptual Framework for AI Capability Evaluations," arXiv preprint arXiv:2506.18213 (2025), https://doi.org/10.48550/arXiv.2506.18213.

[827] Min Zhang, Sato Takumi, Jack Zhang, and Jun Wang, "Case Study: Testing Model Capabilities in Some Reasoning Tasks," arXiv preprint arXiv:2402.09967 (2024), https://doi.org/10.48550/arXiv.2402.09967.

[828] "Resources for Measuring Autonomous AI Capabilities," METR, accessed October 2025, https://metr.org/measuring-autonomous-ai-capabilities/.

[829] Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan, "AI Agents That Matter," *Transactions on Machine Learning Research* (2024), https://openreview.net/forum?id=Zy4uFzMviZ.

[830] O. M. Brown, A. B. Curtis, and J. A. Goodwin, *Principles for Evaluation of AI/ML Model Performance and Robustness* (MIT Lincoln Laboratory, 2021), www.ll.mit.edu/sites/default/files/publication/doc/principles-evaluation-aiml-model-performance-brown-md-62.pdf.

[831] Lalli Myllyaho, Mikko Raatikainen, Tomi Männisto, Tommi Mikkonen, and Jukka K. Nurminen, "Systematic Literature Review of Validation Methods for AI Systems," *Journal of Systems and Software* 181 (2021), https://doi.org/10.1016/j.jss.2021.111050.

[832] Aleksandra Nastoska, Bojana Jancheska, Maryan Rizinski, and Dimitar Trajanov, "Evaluating Trustworthiness in AI: Risks, Metrics, and Applications Across Industries," *Electronics* 14 (2025): 2717, https://doi.org/10.3390/electronics14132717.

[833] Tejal Patwardhan, Rachel Dias, Elizabeth Proehl et al., "GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks," arXiv preprint arXiv:2510.04374 (2025), https://doi.org/10.48550/arXiv.2510.04374.

[834] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, S. C. Hoffman, S. Houde, and K. Kannan, "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," *IBM Journal of Research and Development* 63, no. 4, (2019): 1–15, https://doi.org/10.1147/JRD.2019.2942287.

[835] "Algorithmic Equity Toolkit," ACLU Washington, accessed October 2025, www.aclu-wa.org/AEKit/.

[836] Hossein A. Rahmani, Varsha Ramineni, Emine Yilmaz, Nick Craswell, and Bhaskar Mitra, "Towards Understanding Bias in Synthetic Data for Evaluation," arXiv preprint arXiv:2506.10301 (2025), https://doi.org/10.48550/arXiv.2506.10301.

[837] Joachim Baumann, Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, and Daniele Regoli, "Bias on Demand: A Modelling Framework that Generates Synthetic Data with Bias," in *FAccT '23: 2023 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2023), 1–12, https://doi.org/10.1145/3593013.3594058.

[838] Mina Narayanan, Christian Schoeberl, and Tim G. J. Rudner, *Putting Explainable AI to the Test: A Critical Look at AI Evaluation Approaches* (CSET, February 2025), https://cset.georgetown.edu/publication/putting-explainable-ai-to-the-test-a-critical-look-at-ai-evaluation-approaches/.

[839] Karen Scarfone, Murugiah Souppaya, Amanda Cody, and Angela Orebaugh, *Technical Guide to Information Security Testing and Assessment* (NIST, September 2008), www.nist.gov/privacy-framework/nist-sp-800-115.

[840] "Application Security Testing (AST): Buyers Guide," U.S. General Services Administration, last modified February 23, 2026, www.gsa.gov/technology/it-contract-vehicles-and-purchasing-programs/it-security/application-security-testing.

[841] Paul E. Black, Barbara Guttman, and Vadim Okun, *Guidelines on Minimum Standards for Developer Verification of Software* (NIST, October 2021), https://doi.org/10.6028/NIST.IR.8397.

[842] Xiaogang Zhu, Sheng Wen, Seyit Camtepe, and Yang Xiang, "Fuzzing: A Survey for Roadmap," *ACM Computing Surveys* 54, no. 11s (January 2022), https://doi.org/10.1145/3512345.

[843] Jingquan Ge, Yaowen Zheng, Yuekang Li et al., "OptRCA: A More Efficient and Accurate Approach for Automated Root Cause Analysis and Explanation," *ACM Transactions on Software Engineering and Methodology* (May 2025), https://doi.org/10.1145/3736718.

[844] "Adversarial Testing for Generative AI," Google, accessed September 25, 2025, https://developers.google.com/machine-learning/guides/adv-testing.

[845] Lizhi Lin, Honglin Mu, Zenan Zhai et al., "Against the Achilles' Heel: A Survey on Red Teaming for Generative Models," *Journal of Artificial Intelligence Research* 82 (June 2025), https://doi.org/10.1613/jair.1.17654.

[846] Ken Huang, *Agentic AI Red Teaming Guide* (Cloud Security Alliance, 2025), https://cloudsecurityalliance.org/artifacts/agentic-ai-red-teaming-guide.

[847] Jessica Ji, Vikram Venkatram, and Steph Batalis, "AI Safety Evaluations: An Explainer," CSET, May 28, 2025, https://cset.georgetown.edu/article/ai-safety-evaluations-an-explainer/.

[848] Markov Grey and Charbel-Raphaël Segerie, "Safety by Measurement: A Systematic Literature Review of AI Safety Evaluation Methods," arXiv preprint arXiv:2505.05541 (2025), https://doi.org/10.48550/arXiv.2505.05541.

[849] Zhexin Zhang, Leqi Lei, Lindong Wu et al., "SafetyBench: Evaluating the Safety of Large Language Models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, vol. 1 (Association for Computational Linguistics, 2024), 15537–15553, https://doi.org/10.18653/v1/2024.acl-long.830.

[850] Maribeth Rauh, Nahema Marchal, Arianna Manzini et al., "Gaps in the Safety Evaluation of Generative AI," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (2024): 1200–1217, https://doi.org/10.1609/aies.v7i1.31717.

[851] Miriam Ugarte, Pablo Valle, José Antonio Parejo, Sergio Segura, and Aitor Arrieta, "ASTRAL: A Tool for the Automated Safety Testing of Large Language Models," in *ISSTA Companion '25: Proceedings of the 34th ACM SIGSOFT International Symposium on Software Testing and Analysis* (Association for Computing Machinery, 2025), 31–35, https://doi.org/10.1145/3713081.3731733.

[852] Žana Zekić and Zlatko Stapić, "Software Quality Assessment Standards and Metrics: A Systematic Literature Review," in *Proceedings of the Central European Conference on Information and Intelligent Systems* (CECIIS, 2022), https://archive.ceciis.foi.hr/app/public/conferences/2022/Proceedings/TSE/TSE5.pdf.

[853] Mike Jackson, Steve Crouch, and Bob Baxter, *Software Evaluation: Criteria-Based Assessment*, (Software Sustainability Institute, November 2011), www.software.ac.uk/sites/default/files/SSI-SoftwareEvaluationCriteria.pdf.

[854] Bahar Gezici and Ayça Kolukısa Tarhan, "Systematic Literature Review on Software Quality for AI-Based Software," *Empirical Software Engineering* 27 (2022), https://doi.org/10.1007/s10664-021-10105-2.

[855] Chuanqi Tao, Jerry Gao, and Tiexin Wang, "Testing and Quality Validation for AI Software–Perspectives, Issues, and Practices," *IEEE Access* 7 (2019): 120164–120175, https://doi.org/10.1109/ACCESS.2019.2937107.

[856] Chenyu Wang, Zhou Yang, Ze Shi Li, Daniela Damian, and David Lo, "Quality Assurance for Artificial Intelligence: A Study of Industrial Concerns, Challenges and Best Practices," arXiv preprint arXiv:2402.16391 (2024), https://doi.org/10.48550/arXiv.2402.16391.

[857] Michael Kläs, Rasmus Adler, Lisa Jöckel, Janek Groß, and Jan Reich, "Using Complementary Risk Acceptance Criteria to Structure Assurance Cases for Safety-Critical AI Components," presented at Workshop on Artificial Intelligence Safety, International Joint Conference on Artificial Intelligence, 2021, https://publica.fraunhofer.de/handle/publica/412039.

[858] Ryan Then Ye Tong, Yeow Kai Yuan, Ng Wen Dong, and R. Kanesaraj Ramasamy, "A Review: Methods of Acceptance Testing," in *Proceedings of the International Conference on Technology and Innovation Management* (Atlantis Press, 2022), 76–86, https://doi.org/10.2991/978-94-6463-080-0_7.

[859] Margarida Ferreira, Luis Viegas, Joao Pascoal Faria, and Bruno Lima, "Acceptance Test Generation with Large Language Models: An Industrial Case Study," in *2025 IEEE/ACM International Conference on Automation of Software Test* (IEEE, 2025), 1–11, https://doi.org/10.1109/AST66626.2025.00007.

[860] Ernani César Dos Santos, Patrícia Vilain, and Douglas Hiura Longo, "A Systematic Literature Review to Support the Selection of User Acceptance Testing Techniques," in *ICSE '18: Proceedings of the 40th International Conference on Software Engineering: Companion Proceedings* (Association for Computing Machinery, 2018), 418–419, https://doi.org/10.1145/3183440.3195036.

[861] Muhammad Raees, Inge Meijerink, Ioanna Lykourentzou, Vassilis-Javed Khan, and Konstantinos Papangelis, "From Explainable to Interactive AI: A Literature Review on Current Trends in Human-AI Interaction," *International Journal of Human-Computer Studies* 189 (2024), https://doi.org/10.1016/j.ijhcs.2024.103301.

[862] Lujain Ibrahim, Saffron Huang, Lama Ahmad, Umang Bhatt, and Markus Anderljung, "Towards Interactive Evaluations for Interaction Harms in Human-AI Systems," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, no. 2 (2025), https://ojs.aaai.org/index.php/AIES/issue/view/678.

[863] Tom van Nuenen, Jose Such, and Mark Cote, "Intersectional Experiences of Unfair Treatment Caused by Automated Computational Systems," *Proceedings of the ACM on Human-Computer Interaction* 6, no. 2 (November 2022), https://doi.org/10.1145/3555546.

[864] Xinru Tang, Gabriel Lima, Li Jiang, Lucy Simko, and Yixin Zou, "Beyond 'Vulnerable Populations': A Unified Understanding of Vulnerability from A Socio-Ecological Perspective," *Proceedings of the ACM on Human-Computer Interaction* 9, no. 2 (2025): 1–30, https://doi.org/10.1145/3710935.

[865] Federico Cabitza, Andrea Campagner, Riccardo Angius, Chiara Natali, and Carlo Reverberi, "AI Shall Have No Dominion: On How to Measure Technology Dominance in AI-supported Human Decision-Making," in *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2023), https://doi.org/10.1145/3544548.3581095.

[866] Josh A. Goldstein and Girish Sastry, *How to Assess the Likelihood of Malicious Use of Advanced AI Systems* (CSET, March 2025), https://cset.georgetown.edu/publication/how-to-assess-the-likelihood-of-malicious-use-of-advanced-ai-systems/.

[867] Praveen Damacharla, Ahmad Y. Javaid, Jennie J. Gallimore, and Vijay K. Devabhaktuni, "Common Metrics to Benchmark Human-Machine Teams (HMT): A Review," *IEEE Access* 6 (2018): 38637–38655, https://doi.org/10.1109/ACCESS.2018.2853560.

[868] R. Balachandra, "Critical Signals for Making Go/Nogo Decisions in New Product Development," *Journal of Product Innovation Management* 1, no. 2 (1984): 92–100: https://doi.org/10.1016/S0737-6782(84)80020-X.

[869] Pilar Carbonell-Foulquié, Jose L. Munuera-Alemán, and Ana I. Rodríguez-Escudero, "Criteria Employed for Go/No-Go Decisions When Developing Successful Highly Innovative Products," *Industrial Marketing Management* 33, no. 4 (2004): 307–316, https://doi.org/10.1016/S0019-8501(03)00080-4.

[870] Judith Behrens and Holger Ernst, "What Keeps Managers Away from a Losing Course of Action? Go/Stop Decisions in New Product Development," *Journal of Product Innovation Management* 31, no. 2 (2014): 361–374, https://doi.org/10.1111/jpim.12100.

[871] Emma Kallina and Jatinder Singh, "Stakeholder Involvement for Responsible AI Development: A Process Framework," in *EAAMO '24: Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Association for Computing Machinery, 2024), https://doi.org/10.1145/3689904.3694698.

[872] Elizabeth Bondi, Lily Xu, Diana Acosta-Navas, and Jackson A. Killian, "Envisioning Communities: A Participatory Approach Towards AI for Social Good," in *AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery, 2021), 425–436, https://doi.org/10.1145/3461702.3462612.

[873] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang, "The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice," in *EAAMO '23: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Association for Computing Machinery, 2023), https://doi.org/10.1145/3617694.3623261.

[874] Holden Karnofsky, "If-Then Commitments for AI Risk Reduction," Carnegie Endowment for International Peace, September 23, 2024, https://carnegieendowment.org/research/2024/09/if-then-commitments-for-ai-risk-reduction.

[875] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho, "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance," in *AIES '22: Proceedings of the 2022*

AAAI/ACM Conference on AI, Ethics, and Society (Association for Computing Machinery, 2022), 557–571, https://doi.org/10.1145/3514094.3534181.

[876] Yueqi Li and Sanjay Goel, "Artificial Intelligence Auditability and Auditor Readiness for Auditing Artificial Intelligence Systems," *International Journal of Accounting Information Systems* 56 (2025), https://doi.org/10.1016/j.accinf.2025.100739.

[877] Fred D. Davis, "User Acceptance of Information Technology: System Characteristics, User Perceptions and Behavioral Impacts," *International Journal of Man-Machine Studies* 38, no. 3 (1993): 475–487, https://doi.org/10.1006/imms.1993.1022.

[878] Sage Kelly, Sherrie-Anne Kaye, and Oscar Oviedo-Trespalacios, "What Factors Contribute to the Acceptance of Artificial Intelligence? A Systematic Review," *Telematics and Informatics* 77 (February 2023), https://doi.org/10.1016/j.tele.2022.101925.

[879] E. S. Vorm and David Combs, "Integrating Transparency, Trust, and Acceptance: The Intelligent Systems Technology Acceptance Model (ISTAM)," *International Journal of Human-Computer Interaction* 38, no. 18–20 (2022): 1828–1845, https://doi.org/10.1080/10447318.2022.2070107.

[880] John Winsor, "How to Be Systematic About Adopting AI at Your Company," *Harvard Business Review*, November 22, 2024, https://hbr.org/2024/11/how-to-be-systematic-about-adopting-ai-at-your-company.

[881] Siani Pearson and Tariq Elahi, "Privacy Assurance Checking," in *Digital Privacy: PRIME—Privacy and Identity Management for Europe*, eds. Jan Camenisch, Ronal Leenes, and Dieter Sommer (Springer, 2011), https://doi.org/10.1007/978-3-642-19050-6_16.

[882] Harsh Patel, Dominique Boucher, Emad Fallahzadeh, Ahmed E. Hassan, and Bram Adams, "A State-of-the-Practice Release-Readiness Checklist for Generative AI-Based Software Products: A Gray Literature Survey," *IEEE Software* 42, no. 1 (2025): 74–83, https://doi.org/10.1109/MS.2024.3440190.

[883] Will Henshall, "Nobody Knows How to Safety-Test AI," *Time*, March 21, 2024, https://time.com/6958868/artificial-intelligence-safety-evaluations-risks/.

[884] Michael Feffer, Anusha Sinha, Wesley H. Deng, Zachary C. Lipton, and Hoda Heidari, "Red-Teaming for Generative AI: Silver Bullet or Security Theater?," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (2024): 421–437, https://doi.org/10.1609/aies.v7i1.31647.

[885] Thomas Woodside, "Emergent Abilities in Large Language Models: An Explainer," CSET, April 16, 2024, https://cset.georgetown.edu/article/emergent-abilities-in-large-language-models-an-explainer/.

[886] Jason Wei, Yi Tay, Rishi Bommasani et al., "Emergent Abilities of Large Language Models," *Transactions on Machine Learning Research* (2022), https://openreview.net/forum?id=yzkSU5zdwD.

[887] Peter C. Rigby and Christian Bird, "Convergent Contemporary Software Peer Review Practices," in *ESEC/FSE 2013: Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering* (Association for Computing Machinery, 2013), 202–212, https://doi.org/10.1145/2491411.2491444.

[888] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White et al., "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," in *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Association for Computing, 2020), 33–44, https://doi.org/10.1145/3351095.3372873.

[889] "IEEE Standards for Software Reviews and Audits," IEEE, August 15, 2008, https://doi.org/10.1109/IEEESTD.2008.4601584.

[890] Valerij Dermol and Nada Trunk Širca, "Communication, Company Mission, Organizational Values, and Company Performance," *Procedia—Social and Behavioral Sciences* 238 (2018): 542–551, https://doi.org/10.1016/j.sbspro.2018.04.034.

[891] Matthew Finio and Amanda Downie, "How to Build a Successful AI Strategy," IBM Think, accessed October 2025, www.ibm.com/think/insights/artificial-intelligence-strategy.

[892] Heather E. Canary, Sarah E. Riforgiate, and Yvonne J. Montoya, "The Policy Communication Index: A Theoretically Based Measure of Organizational Policy Communication Practices," *Management Communication Quarterly* 27, no. 4 (2013): 471–502, https://doi.org/10.1177/0893318913494116.

[893] Heather E. Canary, Maria Blevins, and Shireen S. Ghorbani, "Organizational Policy Communication Research: Challenges, Discoveries, and Future Directions," *Communication Reports* 28, no. 1 (2014): 48–64, https://doi.org/10.1080/08934215.2013.865063.

[894] Laurie K. Lewis and Travis L. Russ, "Soliciting and Using Input During Organizational Change Initiatives: What Are Practitioners Doing," *Management Communication Quarterly* 26, no. 2 (2011): 267–294, https://doi.org/10.1177/0893318911431804.

[895] Mustafa C. Ungan, "Standardization Through Process Documentation," *Business Process Management Journal* 12, no. 2 (2006): 135–148, https://doi.org/10.1108/14637150610657495.

[896] *Leveraging COBIT for Effective AI System Governance* (ISACA, 2025), www.isaca.org/resources/white-papers/2025/leveraging-cobit-for-effective-ai-system-governance.

[897] Markus Mykkänen and Kaja Tampere, "Organizational Decision Making: The Luhmannian Decision Communication Perspective," *Journal of Business Studies Quarterly* 5, no. 4 (2014): 131–146, www.proquest.com/scholarly-journals/organizational-decision-making-luhmannian/docview/1542023389/se-2.

[898] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan, "Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies," in *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2023), 1369–1385, https://doi.org/10.1145/3593013.3594087.

[899] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux, "How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies," *Proceedings of the ACM on Human-Computer Interaction* 5, no. 2 (Association for Computing Machinery, 2021): 1–39, https://doi.org/10.1145/3476068.

[900] "Explaining Decisions Made with AI," UK Information Commissioner's Office, accessed October 2025, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/.

[901] "The Executive's AI Playbook," QuantumBlack, McKinsey, accessed October 11, 2025, www.mckinsey.com/capabilities/quantumblack/our-insights/the-executives-ai-playbook.

[902] Gloria Barczak and David Wileman, "Communications Patterns of New Product Development Team Leaders," *IEEE Transactions on Engineering Management* 38, no. 2 (1991): 101–109, https://doi.org/10.1109/17.78406.

[903] Hamzeh Al Amosh and Saleh F. A. Khatib, "Cybersecurity Transparency and Firm Success: Insights from the Australian Landscape," *Australian Economic Papers* 64 (2025): 189–204, https://doi.org/10.1111/1467-8454.12385.

[904] "Reconnaissance," MITRE, last modified April 9, 2025, https://atlas.mitre.org/tactics/AML.TA0002.

[905] Vikram Mohanty, Jude Lim, and Kurt Luther, "What Lies Beneath? Exploring the Impact of Underlying AI Model Updates in AI-Infused Systems," in *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2025), https://doi.org/10.1145/3706598.3713751.

[906] Michael Fagan, Mohammah Maifi Hasan Khan, and Nhan Nguyen, "How Does This Message Make You Feel? A Study of User Perspectives on Software Update/Warning Message Design," *Human-Centric Computing and Information Sciences* 5, no. 36 (2015), https://doi.org/10.1186/s13673-015-0053-y.

[907] Yangheran Piao, Jingjie Li, and Daniel W. Woods, "Measuring the Vulnerability Disclosure Policies of AI Vendors," arXiv preprint arXiv:2509.06136 (2025), https://doi.org/10.48550/arXiv.2509.06136.

[908] Yousra Javed and Ayesha Sajid, "A Systematic Review of Privacy Policy Literature," *ACM Computing Surveys* 57, no. 2 (February 2025), https://doi.org/10.1145/3698393.

[909] Chris Norval, Kristin Cornelius, Jennifer Cobbe, and Jatinder Singh, "Disclosure by Design: Designing Information Disclosures to Support Meaningful Transparency and Accountability," in *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2022), 679–690, https://doi.org/10.1145/3531146.3533133.

[910] "Documentation," UK Information Commissioner's Office, accessed October 26, 2025, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-

decisions-made-with-artificial-intelligence/part-3-what-explaining-ai-means-for-your-organisation/documentation/.

[911] Timnit Gebru, Jamie Morgenstern, Briana Vecchione et al., "Datasheets for Datasets," *Communications of the ACM*, December 1, 2021, https://cacm.acm.org/research/datasheets-for-datasets/.

[912] Bhanu Teja Reddy Maryala, "Data Governance in Generative AI: A Framework for Transparency, Compliance, and Ethical Practice," *Journal of Computer Science and Technology Studies* 7, no. 3 (2025): 964–971, https://doi.org/10.32996/jcsts.2025.7.3.108.

[913] "Statement 13: Establish Data Supply Chain Management Processes," Technical Standard for Government's Use of Artificial Intelligence: Data Statements, Digital.gov.au, Australian Government, accessed October 2025, www.digital.gov.au/policy/ai/AI-technical-standard/ai-technical-standard-statement-13.

[914] "Generative Artificial Intelligence and Open Data: Guidelines and Best Practices," U.S. Department of Commerce, January 16, 2025, www.commerce.gov/news/blog/2025/01/generative-artificial-intelligence-and-open-data-guidelines-and-best-practices.

[915] Dominic Balog-Way, Katherine McComas, and John Besley, "The Evolving Field of Risk Communication," *Risk Analysis* 40 (2020): 2240–2262, https://doi.org/10.1111/risa.13615.

[916] "Guide to Corporate Risk Profiles," Treasury Board of Canada Secretariat, Government of Canada, last modified January 1, 2017, www.canada.ca/en/treasury-board-secretariat/corporate/risk-management/corporate-risk-profiles.html.

[917] Amy Winecoff and Miranda Bogen, "Improving Governance Outcomes Through AI Documentation: Bridging Theory and Practice," in *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2025), https://doi.org/10.1145/3706598.3713814.

[918] Florian Königstorfer and Stefan Thalmann, "AI Documentation: A Path to Accountability," *Journal of Responsible Technology* 11 (2022), https://doi.org/10.1016/j.jrt.2022.100043.

[919] Eli Sherman and Ian Eisenberg, "AI Risk Profiles: A Standards Proposal for Pre-deployment AI Risk Disclosures," *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 21 (2024): 23047–23052, https://doi.org/10.1609/aaai.v38i21.30348.

[920] Noam Kolt, Markus Anderljung, Joslyn Barnhart et al., "Responsible Reporting for Frontier AI Development," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (2024): 768–783, https://doi.org/10.1609/aies.v7i1.31678.

[921] Rishi Bommasani, Kevin Klyman, Shayne Longpre et al., "Foundation Model Transparency Reports," *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (2025): 181–195, https://doi.org/10.1609/aies.v7i1.31628.

[922] Weixin Liang, Nazneen Rajani, Xinyu Yang et al., "Systematic Analysis of 32,111 AI Model Cards Characterizes Documentation Practice in AI," *Nature Machine Intelligence* 6 (2024): 744–753, https://doi.org/10.1038/s42256-024-00857-z.

[923] Jessica Ji, "How to Improve AI Red-Teaming: Challenges and Recommendations," CSET, March 21, 2025, https://cset.georgetown.edu/article/how-to-improve-ai-red-teaming-challenges-and-recommendations/.

[924] Daniela S. Cruzes, Nils B. Moe, and Tore Dybå, "Communication Between Developers and Testers in Distributed Continuous Agile Testing," in *2016 IEEE 11th International Conference on Global Software Engineering* (IEEE, 2016), 59–68, https://doi.org/10.1109/ICGSE.2016.27.

[925] Bran Knowles and John T. Richards, "The Sanction of Authority: Promoting Public Trust in AI," in *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2021): 262–271, https://doi.org/10.1145/3442188.3445890.

[926] Bill Harrison, "Step-by-Step Guide to Managing Your Policy and Procedure Review Process," ComplianceBridge, August 31, 2025, https://compliancebridge.com/policy-and-procedure-review-process/.

[927] "Understanding Continuous Improvement: A Guide for Operational Excellence," Kaizen Institute, accessed October 11, 2025, https://kaizen.com/insights/continuous-improvement-operational-excellence/.

[928] Jagdeep Singh and Harwinder Singh, "Continuous Improvement Philosophy—Literature Review and Directions," *Benchmarking: An International Journal* 22, no. 1 (February 2, 2015): 75–119, https://doi.org/10.1108/BIJ-06-2012-0038.

[929] *The IIA's Artificial Intelligence Auditing Framework* (Institute of Internal Auditors, 2023), www.theiia.org/en/content/tools/professional/2023/the-iias-updated-ai-auditing-framework/.

[930] *Auditing Artificial Intelligence* (ISACA, 2018), www.isaca.org/resources/white-papers/auditing-artificial-intelligence.

[931] Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji, "Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling," in *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2025), https://doi.org/10.1145/3706598.3713301.

[932] Alberto-Tomas Delso-Vicente, Luis Diaz-Marcos, Oscar Aguado-Tevar, and María García de Blanes-Sebastián, "Factors Influencing Employee Compliance with Information Security Policies: A Systematic Literature Review of Behavioral and Technological Aspects in Cybersecurity," *Future Business Journal* 11, no. 28 (2025), https://doi.org/10.1186/s43093-025-00452-7.

[933] Yenny Farinas Diaz and Marc L. Resnick, "A Model to Predict Employee Compliance with Employee Corporate's Safety Regulations Factoring Risk Perception," *Proceedings of the Human Factors*

and Ergonomics Society Annual Meeting 44, no. 27 (2000): 323–326, https://doi.org/10.1177/154193120004402711.

[934] Betsy Macht and Anne Davis, "Strategies to Influence a Quality and Compliance Culture," *International Journal of Applied Management and Technology* 17, no. 1 (2018): 68–82, https://doi.org/10.5590/IJAMT.2018.17.1.06.

[935] Ralph M. Foorthuis, "Tactics for Internal Compliance: A Literature Review," arXiv preprint arXiv:2008.03775 (2020), https://doi.org/10.48550/arXiv.2008.03775.

[936] Kirstie Ball, *Electronic Monitoring and Surveillance in the Workplace: Literature Review and Policy Recommendations* (Publications Office of the European Union, 2021), https://doi.org/10.2760/451453.

[937] Teshan S. Bunwaree, Katarzyna Stawarz, Philippa Collins, and Sandy J. J. Gould, "Boss Is aWare—Are You? Employee Comprehension and Legal Awareness of Workplace Monitoring," in *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2025), https://doi.org/10.1145/3706598.3713651.

[938] Angie Zhang and Min Kyung Lee, "Knowledge Workers' Perspectives on AI Training for Responsible AI Use," in *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2025), https://doi.org/10.1145/3706598.3714100.

[939] Jaromir Savelka, Can Kultur, Arav Agarwal et al., "AI Technicians: Developing Rapid Occupational Training Methods for a Competitive AI Workforce," arXiv preprint arXiv:2501.10579 (2025), https://doi.org/10.48550/arXiv.2501.10579.

[940] Jayarajan Samuel, Sridhar Nerur, RadhaKanta Mahapatra, and Brian White, "Building AI Talent in Organizations—An Experiential Learning Approach," *Journal of Information Systems Education* 36, no. 3 (2025): 277–286, https://doi.org/10.62273/NRQW1204.

[941] *Bridging the AI Skills Gap: Is Training Keeping Up?* (OECD, April 24, 2025), www.oecd.org/en/publications/bridging-the-ai-skills-gap_66d0702e-en.html.

[942] *What's Working? Navigating the AI Revolution and the Shifting Future of Work*" (Adecco Group, 2023), www.adeccogroup.com/global-workforce-of-the-future-research-2023.

[943] Hannah Mayer, Lareina Yee, Michael Chui, and Roger Roberts, *Superagency in the Workplace: Empowering People to Unlock AI's Full Potential* (McKinsey, January 28, 2025), www.mckinsey.com/capabilities/mckinsey-digital/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work.

[944] Muhammad Mudassar Yamin, Basel Katt, and Mariusz Nowostawski, "Serious Games as a Tool to Model Attack and Defense Scenarios for Cyber-Security Exercises," *Computers & Security* 110 (2021), https://doi.org/10.1016/j.cose.2021.102450.

[945] Ryan Hofschneider (ryanhofdotgov), "18F / development-guide / incident-response-drills.md," GitHub, accessed October 17, 2025, https://github.com/18F/development-guide/blob/main/_pages/security/incident-response-drills.md.

[946] Ashley O'Neill, Sean B. Maynard, Atif Ahmad, and Justin Filippou, "Cybersecurity Incident Response in Organisations: A Meta-level Framework for Scenario-Based Training," *ACIS 2022 Proceedings* 35 (2022), https://aisel.aisnet.org/acis2022/35.

[947] Alexander E. Grojek, Leslie F. Sikos, David J. Holmes, and Oliver Guidetti, "Incident Response Drills on Cyber Ranges," in *Psybersecurity: Human Factors of Cyber Defence*, eds. Oliver Guidetti, Mohiuddin Ahmed, Craig Speelman (CRC Press, 2024), 202–226, https://doi.org/10.1201/9781032664859.

[948] Robert Sheldon and Gavin Wright, "EAI (Enterprise Application Integration)," TechTarget, March 5, 2024, www.techtarget.com/searchapparchitecture/definition/EAI-enterprise-application-integration.

[949] Muhammad Ilyas, Siffat Ullah Khan, and Nasir Rashid, "Empirical Validation of Software Integration Practices in Global Software Development," *SN Computer Science* 1 (2020), https://doi.org/10.1007/s42979-020-00175-2.

[950] Muhammad Ilyas, Siffat Ullah Khan, Habib Ullah Khan, and Nasir Rashid, "Software Integration Model: An Assessment Tool for Global Software Development Vendors," *Journal of Software Evolution and Process* 36, no. 4 (April 2024), https://doi.org/10.1002/smr.2540.

[951] Yang Lu, Rabimba Karanjai, Dana Alsagheer et al., "LogBabylon: A Unified Framework for Cross-Log File Integration and Analysis," in *SAC '25: Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing* (Association for Computing Machinery, 2025), 1953–1960, https://doi.org/10.1145/3672608.3707883.

[952] Bainian Wu, Shuo Zhang, Yaping Liu, and Zhikai Yang, "A Survey of Network Asset Detection Technology," in *Network Simulation and Evaluation* (Springer, 2024), https://doi.org/10.1007/978-981-97-4522-7_1.

[953] Igor Kotenko, Elena Doynikova, Andrey Fedorchenko, and Vasily Desnitsky, "Automation of Asset Inventory for Cyber Security: Investigation of Event Correlation-Based Technique," *Electronics* 11, no. 15 (2022): 2368, https://doi.org/10.3390/electronics11152368.

[954] Alexandra Wood, Micah Altman, Aaron Bembenek et al., "Differential Privacy: A Primer for a Non-technical Audience," *Vanderbilt Journal of Entertainment and Technology Law* 21, no. 1 (2018): 209–276, https://scholarship.law.vanderbilt.edu/jetlaw/vol21/iss1/4.

[955] Joseph P. Near and Chiké Abuah, *Programming Differential Privacy*, vol. 1 (self-published, 2021), https://programming-dp.com/.

[956] Arvind Narayanan and Vitaly Shmatikov, "Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)," arXiv preprint arXiv:cs/0610105 (2006), https://doi.org/10.48550/arXiv.cs/0610105.

[957] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye, "Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models," *Nature Communications* 10 (2019), https://doi.org/10.1038/s41467-019-10933-3.

[958] "How Do We Ensure Lawfulness in AI?," UK Information Commissioner's Office, October 28, 2024, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-lawfulness-in-ai/.

[959] Mindy Nunez Duffourc, Sara Gerke, and Konrad Kollnig, "Privacy of Personal Data in the Generative AI Data Lifecycle," *Journal of Intellectual Property and Entertainment Law* 13, no. 2 (July 8, 2024), https://jipel.law.nyu.edu/privacy-of-personal-data-in-the-generative-ai-data-lifecycle/.

[960] "PAI's Guidance for Safe Foundation Model Deployment: A Framework for Collective Action," Partnership on AI, last modified July 2024, https://partnershiponai.org/modeldeployment/.

[961] Mojtaba Shahin, Muhammad Ali Babar, and Liming Zhu, "Continuous Integration, Delivery and Deployment: A Systematic Review on Approaches, Tools, Challenges and Practices," *IEEE Access* 5 (2017): 3909–3943, https://doi.org/10.1109/ACCESS.2017.2685629.

[962] *Enforce Secure Automated Deployment Practices through Infrastructure as Code* (NSA, March 2024), https://media.defense.gov/2024/Mar/07/2003407857/-1/-1/0/CSI-CloudTop10-Infrastructure-as-Code.PDF.

[963] NSA, CISA, FBI et al., *Deploying AI Systems Securely: Best Practices for Deploying Secure and Resilient AI Systems* (NSA, April 2024), https://media.defense.gov/2024/Apr/15/2003439257/-1/-1/0/CSI-DEPLOYING-AI-SYSTEMS-SECURELY.PDF.

[964] Cameron F. Kerry, Joshua P. Meltzer, Andrea Renda, Alex Engler, and Rosanna Fanni, *Strengthening International Cooperation on AI: Progress Report* (Brookings, October 25, 2021), www.brookings.edu/articles/strengthening-international-cooperation-on-ai/.

[965] Hadrien Pouget, Claire Dennis, Jon Bateman et al., *The Future of International Scientific Assessments of AI's Risks* (Carnegie Endowment for International Peace, August 2024), https://carnegieendowment.org/research/2024/08/the-future-of-international-scientific-assessments-of-ais-risks.

[966] Bureau of Arms Control and Nonproliferation, "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," U.S. Department of State, November 9, 2023, www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/.

[967] Emelia Probasco, Matthew Burtell, Helen Toner, and Tim G. J. Rudner, "Not Oracles of the Battlefield: Safety Considerations for AI-Based Military Decision Support Systems," in *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society* (AAAI Press, 2024), 1157–1165, https://doi.org/10.1609/aies.v7i1.31712.

[968] Zoe Stanley-Lockman, *Responsible and Ethical Military AI: Allies and Allied Perspectives* (CSET, August 2021), https://doi.org/10.51593/20200091.

[969] Joint Cyber Defense Collaborative, *JCDC AI Cybersecurity Collaboration Playbook* (CISA, January 14, 2025), www.cisa.gov/resources-tools/resources/ai-cybersecurity-collaboration-playbook.

[970] Ellen P. Goodman, *Artificial Intelligence Accountability Policy Report* (National Telecommunications and Information Administration, March 2024), www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report.

[971] Ren Bin Lee Dixon and Heather Frase, *AI Incidents: Key Components for a Mandatory Reporting Regime* (CSET, January 2025), https://cset.georgetown.edu/publication/ai-incidents-key-components-for-a-mandatory-reporting-regime/.

[972] John Croxton, David Robusto, Satya Thallam, and Doug Calidas, "Message Incoming: Establish an AI Incident Reporting System," Federation of American Scientists, June 25, 2024, https://fas.org/publication/establishing-an-ai-incident-reporting-system/.

[973] Renée Sieber, Ana Brandusescu, Abigail Adu-Daako, and Suthee Sangiambut, "Who Are the Publics Engaging in AI?," *Public Understanding of Science* 33, no. 5 (2024): 634–653, https://doi.org/10.1177/09636625231219853.

[974] Gene Rowe and Lynn J. Frewer, "A Typology of Public Engagement Mechanisms," *Science, Technology, & Human Values* 30, no. 2 (2005): 251–190, https://doi.org/10.1177/0162243904271724.

[975] Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, Didar Zowghi, and Aurelie Jacquet, "Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering," *ACM Computing Surveys* 56, no. 7 (July 2024), https://doi.org/10.1145/3626234.

[976] *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space* (Forum on Information and Democracy, February 2024), https://informationdemocracy.org/publications/artificial-intelligence-as-a-public-good-ensuring-democratic-control-of-ai-in-the-information-space/.

[977] Benjamin S. Bucknall and Robert F. Trager, *Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers' Model Access Requirements* (Centre for the Governance of AI, October 2023), www.governance.ai/research-paper/structured-access-for-third-party-research-on-frontier-ai-models.

[978] Andrej Krištofík, Jakub Vostoupal, Kamil Malinka, František Kasl, and Pavel Loutocký, "Beyond the Bugs: Enhancing Bug Bounty Programs Through Academic Partnerships," in *ARES '24: Proceedings of the 19th International Conference on Availability, Reliability and Security* (Association for Computing Machinery, 2024), https://doi.org/10.1145/3664476.3670455.

[979] Noemi Dreksler, Harry Law, Chloe Ahn et al., *What Does the Public Think About AI? An Overview of the Public's Attitudes Towards AI and a Resource for Future Research* (Centre for the Governance of AI, January 22, 2025), www.governance.ai/research-paper/what-does-the-public-think-about-ai.

[980] Colleen McClain, Brian Kennedy, Jeffrey Gottfried, Monica Anderson, and Giancarlo Pasquini, "How the U.S. Public and AI Experts View Artificial Intelligence," Pew Research Center, April 3, 2025, www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/.

[981] *A Blueprint for Equity and Inclusion in Artificial Intelligence* (World Economic Forum, June 2022), www.weforum.org/publications/a-blueprint-for-equity-and-inclusion-in-artificial-intelligence/.

[982] Leah Hope Ajmani, Nuredin Ali Abdelkadir, and Stevie Chancellor, "Secondary Stakeholders in AI: Fighting for, Brokering, and Navigating Agency," in *FAccT '25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2025), 1095–1107, https://doi.org/10.1145/3715275.3732071.

[983] *Disconnected AI: The Unmet Expectations of Consumers and Workers* (UTS Human Technology Institute, March 2025), www.uts.edu.au/globalassets/sites/default/files/2025-03/HTI-Insight-Summary-Disconnected-AI.pdf.

[984] Marjorie Kinney, Maria Anastasiadou, Mijail Naranjo-Zolotov, and Vitor Santos, "Expectation Management in AI: A Framework for Understanding Stakeholder Trust and Acceptance of Artificial Intelligence Systems," *Heliyon* 10, no. 7 (2024), https://doi.org/10.1016/j.heliyon.2024.e28562.

[985] Meredith Caldwell, Niels Wouters, Llewellyn Spink, and Nicholas Davis, *From Invisible to Involved: A Guide to Worker Engagement on AI* (UTS Human Technology Institute, June 2025), www.uts.edu.au/globalassets/sites/default/files/2025-06/25.06.04-hti-guide-to-worker-engagement-on-ai.pdf.

[986] "Compliance Programs: Employee Reporting Mechanisms," *Practical Law*, May 2024, www.reuters.com/practical-law-the-journal/transactional/compliance-programs-employee-reporting-mechanisms-2024-05-01/.

[987] Barbara Culiberg, and Katarina Katja Mihelič, "The Evolution of Whistleblowing Studies: A Critical Review and Research Agenda," *Journal of Business Ethics* 146 (2107): 787–803, https://doi.org/10.1007/s10551-016-3237-0.

[988] Ravinithesh Annapureddy, Alessandro Fornaroli, and Daniel Gatica-Perez, "Generative AI Literacy: Twelve Defining Competencies," *Digital Government: Research and Practice* 6, no. 1 (March 2025), https://doi.org/10.1145/3685680.

[989] Emanuel Rieder, "Strategic Alignment and AI Adoption: How Organizational Factors Shape Perceived Productivity," *International Journal of Management and Accounting* 7, no. 4 (August 20, 2025), https://doi.org/10.34104/ijma.025.01840192.

[990] "How to Manage Your Security When Engaging a Managed Service Provider," Australian Signals Directorate, October 6, 2021, www.cyber.gov.au/business-government/supplier-cyber-risk-management/managed-service-providers/how-to-manage-your-security-when-engaging-a-managed-service-provider.

[991] Theresa Sobb, Benjamin Turnbull, and Nour Moustafa, "Supply Chain 4.0: A Survey of Cyber Security Challenges, Solutions and Future Directions," *Electronics* 9, no. 11 (2020), https://doi.org/10.3390/electronics9111864.

[992] Jennifer Cobbe, Michael Veale, and Jatinder Singh, "Understanding Accountability in Algorithmic Supply Chains," in *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2023), 1186–1197, https://doi.org/10.1145/3593013.3594073.

[993] *Complaint Mechanisms: Reference Guide for Good Practice* (Transparency International, 2016), https://knowledgehub.transparency.org/product/complaint-mechanisms-reference-guide-for-good-practice.

[994] *Grievance Mechanism Toolkit* (Compliance Advisor Ombudsman, 2016), www.cao-ombudsman.org/grm/purpose-design-and-implementation.html.

[995] "The Right to Object to the Use of Your Data," UK Information Commissioner's Office, accessed October 17, 2025, https://ico.org.uk/for-the-public/the-right-to-object-to-the-use-of-your-data/.

[996] Hana Habib, Yixin Zou, Aditi Jannu et al., "An Empirical Analysis of Data Deletion and Opt-Out Choices on 150 Websites," in *Fifteenth Symposium on Usable Privacy and Security* (USENIX Association, 2019), 387–406, www.usenix.org/conference/soups2019/presentation/habib.

[997] Deirdre K. Mulligan, Daniel Kluttz, and Nitin Kohli, "Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions," SSRN, July 7, 2019, http://dx.doi.org/10.2139/ssrn.3311894.

[998] Mireia Yurrita, Himanshu Verma, Agathe Balayn, Kars Alfrink, Ujwal Gadiraju, and Alessandro Bozzon, "Identifying Algorithmic Decision Subjects' Needs for Meaningful Contestability," *Proceedings of the ACM Human-Computer Interaction* 9, no. 7 (November 2025): 1–29, https://doi.org/10.1145/3757415.

[999] *A Plan for Global Engagement on AI Standards* (NIST, July 2024), https://doi.org/10.6028/NIST.AI.100-5.

[1000] Jonathon Monken, Fernando Maymi, Dan Bennet et al., *Cyber Mutual Assistance Workshop Report* (Software Engineering Institute, Carnegie Mellon University, February 2018), www.sei.cmu.edu/library/cyber-mutual-assistance-workshop-report/.

[1001] *The ESCC's Cyber Mutal Assistance Program* (Electricity Subsector Coordinating Council, August 2024), www.electricitysubsector.org/CMA/.

[1002] Ratun Rahman, "Federated Learning: A Survey on Privacy-Preserving Collaborative Intelligence," arXiv preprint arXiv:2504.17703 (2025), https://doi.org/10.48550/arXiv.2504.17703.

[1003] Saurabh Gupta, Robert P. Bostrom, and Mark Huber, "End-User Training Methods: What We Know, Need to Know," *ACM SIGMIS Database* 41, no. 4 (November 2010): 9–39, https://doi.org/10.1145/1899639.1899641.

[1004] Simone Stumpf, Vidya Rajaram, Lida Li et al., "Toward Harnessing User Feedback for Machine Learning," in *IUI '07: Proceedings of the 12th International Conference on Intelligent User Interfaces* (Association for Computing Machinery, 2007), 82–91, https://doi.org/10.1145/1216295.1216316.

[1005] Wesley Hanwen Deng, Wang Claire, Howard Ziyu Han, Jason I. Hong, Kenneth Holstein, and Motahhare Eslami, "WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI," *Proceedings of the ACM on Human-Computer Interaction* 9, no. 7 (November 2025): 1–35, https://doi.org/10.1145/3757702.

[1006] Kazuo Okamura and Seiji Yamada, "Adaptive Trust Calibration for Human-AI Collaboration," *PLOS One* 15, no. 2 (2020), https://doi.org/10.1371/journal.pone.0229132.

[1007] Lauren Kahn, Emelia Probasco, and Ronnie Kinoshita, *AI Safety and Automation Bias: The Downside of Human-in-the-Loop* (CSET, November 2024), https://cset.georgetown.edu/publication/ai-safety-and-automation-bias/.

[1008] Virginija Ramašauskienė, Erika Župerkienė, and Ligita Šimanskienė, "Situational Awareness in Leadership: Application of Methods in Business Organisations," *Administrative Sciences* 15, no. 6 (2025), https://doi.org/10.3390/admsci15060210.

[1009] Julia Brock and James Andrew Lewis, "Criteria for Cyber Situational Awareness," Center for Strategic International Studies, May 22, 2025, www.csis.org/analysis/criteria-cyber-situational-awareness.

[1010] "Situational Awareness," Software Engineering Institute, Carnegie Mellon University, accessed October 2025, www.sei.cmu.edu/situational-awareness/.

[1011] Matteo Pedrini and Laura Maria Ferri, "Stakeholder Management: A Systematic Literature Review," *Corporate Governance* 19, no. 1 (2019): 44–59, https://doi.org/10.1108/CG-08-2017-0172.

[1012] Blair Epstein, Julia McClatchy, Kurt Strovink, and Eric Sherman, "How the Best CEOs Build Lasting Stakeholder Relationships," McKinsey, November 26, 2024, www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/how-the-best-ceos-build-lasting-stakeholder-relationships.

[1013] Douglas M. Lambert and Matthew A. Schwieterman, "Supplier Relationship Management as a Macro Business Process," *Supply Chain Management: An International Journal* 17, no. 3 (April 27, 2012): 337–352. https://doi.org/10.1108/13598541211227153.

[1014] *AI Accountability Policy Report* (National Telecommunications and Information Administration, March 27, 2024), www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report.

[1015] Lena Theodora Schramm, Raymund Lin, and Hsiao-Lan Wei, "Beyond Technical Training: A Cybersecurity Skills Framework for Non-professionals," in *SIGMIS-CPR '25: Proceedings of the 2025 Computers and People Research Conference* (Association for Computing Machinery, 2025), https://doi.org/10.1145/3716489.3728444.

[1016] Jonah Stegman, Patrick J. Trottier, Caroline Hillier, Hassan Khan, and Mohammad Mannan, "My Privacy for their Security: Employees' Privacy Perspectives and Expectations when Using Enterprise Security Software," in *32nd USENIX Security Symposium* (USENIX Association, 2023), 3583–3600, www.usenix.org/conference/usenixsecurity23/presentation/stegman.

[1017] Duri Long and Brian Magerko. "What Is AI Literacy? Competencies and Design Considerations," in *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2020), 1–16, https://doi.org/10.1145/3313831.3376727.

[1018] Fengchun Miao, Kelly Shiohira, and Natalie Lao, *AI Competency Framework for Students* (UNESCO, August 2024), www.unesco.org/en/articles/ai-competency-framework-students.

[1019] Nada R. Sanders and John D. Wood, "The Skills Your Employees Need to Work Effectively with AI," *Harvard Business Review*, November 3, 2023, https://hbr.org/2023/11/the-skills-your-employees-need-to-work-effectively-with-ai.

[1020] Marc Schmitt and Ivan Flechais, "Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing," *Artificial Intelligence Review* 57 (2024), https://doi.org/10.1007/s10462-024-10973-2.

[1021] Katerina Sedova, Christine McNeill, Aurora Johnson, Aditi Joshi, and Ido Wulkan, *AI and the Future of Disinformation Campaigns* (CSET, December 2021), https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns-2/.

[1022] Mica R. Endsley, "Supporting Human-AI Teams: Transparency, Explainability, and Situation Awareness," *Computers in Human Behavior* 140 (2023), https://doi.org/10.1016/j.chb.2022.107574.

[1023] Chief Data Office, "AI System Inventory Guidance Document," New York State Office of Information Technology Services, March 4, 2025, https://its.ny.gov/ai-inventory-guidance.

[1024] Lukas Rauh, Mel-Rick Süner, Daniel Schel, and Thomas Bauernhansl, "AI Asset Management for Manufacturing (AIM4M): Development of a Process Model for Operationalization," arXiv preprint arXiv:2509.11691 (2025), https://doi.org/10.48550/arXiv.2509.11691.

[1025] Thomas Baumer, Tobias Reittinger, Sascha Kern, and Günther Pernul, "Digital Nudges for Access Reviews: Guiding Deciders to Revoke Excessive Authorizations," in *SOUPS 2024: Twentieth Symposium on Usable Privacy and Security* (USENIX Association, 2024), 239–258, www.usenix.org/conference/soups2024/presentation/baumer.

[1026] Mathieu Gorge, "Making Sense of Log Management for Security Purposes—An Approach to Best Practice Log Collection, Analysis and Management," *Computer Fraud and Security* 2007, no. 5 (2007), https://doi.org/10.1016/S1361-3723(07)70047-7.

[1027] Lujo Bauer, Lorrie Faith Cranor, Robert W. Reeder, Michael K. Reiter, and Kami Vaniea, "Real Life Challenges in Access-Control Management," in *CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2009), 899–908, https://doi.org/10.1145/1518701.1518838.

[1028] "IT Security Procedural Guides," U.S. General Services Administration, accessed October 2025, www.gsa.gov/policy-regulations/policy/information-technology-policy/it-security-procedural-guides.

[1029] Abhishek Goyal, "Optimising Software Lifecycle Management Through Predictive Maintenance: Insights and Best Practices," *International Journal of Science and Research Archive* 7, no. 2 (2022): 693–702, https://doi.org/10.30574/ijsra.2022.7.2.0348.

[1030] Henry George and Austin Arnett, "Implementing Cybersecurity Best Practices for Electrical Infrastructure in a Refinery: A Case Study," *IEEE Industry Applications Magazine* 27 (2021): 18–24, https://doi.org/10.1109/MIAS.2021.3063095.

[1031] Ben Tailor-Hamblin, "Best Practices for Retiring Applications Before Decommissioning Infrastructure," Amazon Web Services, January 2023, https://docs.aws.amazon.com/prescriptive-guidance/latest/migration-app-retirement-best-practices/welcome.html.

[1032] "Decommissioning Assets," UK National Cyber Security Centre, May 20, 2025, www.ncsc.gov.uk/guidance/decommissioning-assets.

[1033] Patrick Murmann and Simone Fischer-Hübner, "Tools for Achieving Usable Ex Post Transparency: A Survey," *IEEE Access* 5 (2017): 22965–22991, https://doi.org/10.1109/ACCESS.2017.2765539.

[1034] "Right of Access," UK Information Commissioner's Office, accessed October 17, 2025, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/right-of-access/.

[1035] Sanjam Garg, Shafi Goldwasser, and Prashant Nalini Vasudevan, "Formalizing Data Deletion in the Context of the Right to Be Forgotten," in *Advances in Cryptology—EUROCRYPT 2020* (Springer, 2020), https://doi.org/10.1007/978-3-030-45724-2_13.

[1036] Ben Wolford, "Everything You Need to Know About the Right to Be Forgotten," GDPR.eu, accessed September 2025, https://gdpr.eu/right-to-be-forgotten.

[1037] "Data Protection Laws of the World," DLA Piper, accessed October 17, 2025, www.dlapiperdataprotection.com/.

[1038] Kentaro Hoffman, Stephen Salerno, Jeff Leek, and Tyler McCormick, "Some Models Are Useful, but for How Long?: A Decision Theoretic Approach to Choosing When to Refit Large-Scale Prediction Models," arXiv preprint arXiv:2405.13926 (2024), https://doi.org/10.48550/arXiv.2405.13926.

[1039] Sijia Liu, Yuanshun Yao, Jingham Jia et al., "Rethinking Machine Unlearning for Large Language Models," *Nature Machine Intelligence* 7 (2025): 181–194, https://doi.org/10.1038/s42256-025-00985-0.

[1040] Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid, "The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities," arXiv preprint arXiv:2408.13296 (2024), https://doi.org/10.48550/arXiv.2408.13296.

[1041] Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li, "Knowledge Editing for Large Language Models: A Survey," *ACM Computing Surveys* 57, no. 3 (March 2025), https://doi.org/10.1145/3698590.

[1042] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang, "Automatically Correcting Large Language Models: Surveying the Landscape of Diverse Automated Correction Strategies," *Transactions of the Association for Computational Linguistics* 12 (2024): 484–506, https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00660/120911.

[1043] Surya Gangadhar Patchipala, "Tackling Data and Model Drift in AI: Strategies for Maintaining Accuracy During ML Model Inference," *International Journal of Science and Research Archive* 10, no. 2 (2023): 1198–1209, https://doi.org/10.30574/ijsra.2023.10.2.0855.

[1044] "Guidance," Vulnerability Management, UK National Cyber Security Centre, February 12, 2024, www.ncsc.gov.uk/collection/vulnerability-management/guidance.

[1045] "Known Exploited Vulnerabilities Catalog," CISA, accessed September 2025, www.cisa.gov/known-exploited-vulnerabilities-catalog.

[1046] "AI Vulnerability Database," AVID, accessed September 2025, https://avidml.org/.

[1047] "What Are the Risks from Artificial Intelligence?," MIT AI Risk Initiative, accessed September 2025, https://airisk.mit.edu/.

[1048] "Vulnerability Scanning Tools and Services," UK National Cyber Security Centre, January 19, 2021, www.ncsc.gov.uk/guidance/vulnerability-scanning-tools-and-services.

[1049] "OWASP Vulnerability Management Guide," OWASP, accessed September 2025, https://owasp.org/www-project-vulnerability-management-guide/.

[1050] "Patching and Updating," SAMM, OWASP, accessed September 2025, https://owaspsamm.org/model/operations/environment-management/stream-b/.

[1051] Spiros Alexiou, "Practical Patch Management and Mitigation," ISACA, May 1, 2019, www.isaca.org/resources/isaca-journal/issues/2019/volume-3/practical-patch-management-and-mitigation.

[1052] Marcus Comiter, *Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It* (Harvard Kennedy School Belfer Center for Science and International Affairs, August 2019), www.belfercenter.org/publication/AttackingAI.

[1053] Andrew Lohn and Wyatt Hoffman, *Securing AI: How Traditional Vulnerability Disclosure Must Adapt* (CSET, March 2022), https://cset.georgetown.edu/publication/securing-ai-how-traditional-vulnerability-disclosure-must-adapt/.

[1054] Esmat Zaidan and Imad Antoine Ibrahim, "AI Governance in a Complex and Rapidly Changing Regulatory Landscape: A Global Perspective," *Humanities and Social Sciences Communications* 11 (2024), https://doi.org/10.1057/s41599-024-03560-x.

[1055] Naveen Sukumaran, "Evolving Landscape of AI Regulation: Global Trends and National Responses," in *Artificial Intelligence and Law* (Satyam Law International, 2025), 26–64, https://dx.doi.org/10.2139/ssrn.5356651.

[1056] Adebola Folorunso, Ifeoluwa Wada, Bunmi Samuel, and Viqaruddin Mohammed, "Security Compliance and Its Implication for Cybersecurity," *World Journal of Advanced Research and Reviews* 24, no. 1 (2024), https://doi.org/10.30574/wjarr.2024.24.1.3170.

[1057] Sri Nikhil Gupta Gourisetti, Michael Mylrea, Hayden Reeve, Julia A. Rotondo, Grant T. Richards, and Jacob Irwin, *Facility Cybersecurity Framework Best Practices*, version 2.0 (Pacific Northwest National Laboratory, October 2021), https://www.osti.gov/servlets/purl/1829733.

[1058] O. Kruzhilko, O. Polukarov, S. Vambol et al., "Control of the Workplace Environment by Physical Factors and SMART Monitoring," *International Scientific Journal: Archives of Materials Science and Engineering* 103, no. 1 (2020): 18–29, https://doi.org/10.5604/01.3001.0014.1770.

[1059] "Implementing SIEM and SOAR Platforms: Practitioner Guidance," Australian Signals Directorate, May 27, 2025, www.cyber.gov.au/business-government/detecting-responding-to-threats/event-logging/implementing-siem-soar-platforms/implementing-siem-and-soar-platforms-practitioner-guidance.

[1060] "Priority Logs for SIEM Ingestion: Practitioner Guidance," Australian Signals Directorate, May 27, 2025, www.cyber.gov.au/business-government/detecting-responding-to-threats/event-logging/implementing-siem-soar-platforms/priority-logs-for-siem-ingestion-practitioner-guidance.

[1061] Jan Svacina, Jackson Raffety, Connor Woodahl et al., "On Vulnerability and Security Log Analysis: A Systematic Literature Review on Recent Trends," in *RACS '20: Proceedings of the International Conference on Research in Adaptive and Convergent Systems* (Association for Computing Machinery, 2020), 175–180, https://doi.org/10.1145/3400286.3418261.

[1062] Nan Sun, Jun Zhang, Paul Rimba, Shang Gao, Leo Yu Zhang, and Yang Xiang, "Data-Driven Cybersecurity Incident Prediction: A Survey," *IEEE Communications Surveys & Tutorials* 21, no. 2 (2019): 1744–1772, https://doi.org/10.1109/COMST.2018.2885561.

[1063] Chris Johnson, Lee Badger, David Waltermire, Julie Snyder, and Clem Skorupka, *Guide to Cyber Threat Information Sharing* (NIST, October 2016), http://dx.doi.org/10.6028/NIST.SP.800-150.

[1064] Thomas D. Wagner, Khaled Mahbub, Esther Palomar, and Ali E. Abdallah, "Cyber Threat Intelligence Sharing: Survey and Research Directions," *Computers & Security* 87 (2019), https://doi.org/10.1016/j.cose.2019.101589.

[1065] Ali Pala and Jun Zhuang, "Information Sharing in Cybersecurity: A Review," *Decision Analysis* 16, no. 3 (2019): 172–196, https://doi.org/10.1287/deca.2018.0387.

[1066] "About ISACs," National Council of ISACs, accessed October 17, 2025, www.nationalisacs.org/about-isacs.

[1067] Simone Busetti and Francesco Maria Scanni, "Evaluating Incident Reporting in Cybersecurity. From Threat Detection to Policy Learning," *Government Information Quarterly* 42, no. 1 (2025), https://doi.org/10.1016/j.giq.2024.102000.

[1068] "Types of Access You May Provide to a Third Party," Captain Compliance, May 25, 2025, https://captaincompliance.com/education/types-of-access-you-may-provide-to-a-third-party/.

[1069] *The State of Cybersecurity and Third-Party Remote Access Risk* (Imprivata, 2024), https://security.imprivata.com/wp-state-of-cybersecurity-third-party-remote-access-register.html.

[1070] Jodi L. Short, Michael W. Toffel, and Andrea R. Hugill, "Monitoring Global Supply Chains," *Strategic Management Journal* 37, no. 9 (2016): 1878–1897, https://doi.org/10.1002/smj.2417.

[1071] Kyle Chin, "Ongoing Monitoring for Third-Party Risk Management (Full Guide)," UpGuard, June 25, 2025, www.upguard.com/blog/ongoing-monitoring-for-tprm.

[1072] "Sharing Information to Get Ahead of Supply Chain Risks," CISA, last modified October 25, 2021, www.cisa.gov/news-events/news/sharing-information-get-ahead-supply-chain-risks.

[1073] Lu Xu, Yanhui Li, Yanwei Lin, Chaofeng Tang, and Qi Yao, "Supply Chain Cybersecurity Investments with Interdependent Risks Under Different Information Exchange Modes," *International Journal of Production Research* 62, no. 6 (2023): 2034–2059, https://doi.org/10.1080/00207543.2023.2206923.

[1074] Abel Yeboah-Ofori, Shareeful Islam, Sin Wee Lee, Khan Muhammad, and Meteb Altaf, "Cyber Threat Predictive Analytics for Improving Cyber Supply Chain Security," *IEEE Access* 9 (2021): 94318–94337, https://doi.org/10.1109/ACCESS.2021.3087109.

[1075] Georgia Killcrece, Klaus-Peter Kossakowski, Robin Ruefle, and Mark Zajicek, *Organizational Models for Computer Security Incident Response Teams (CSIRTs)* (Software Engineering Institute, Carnegie Mellon University, December 2003), www.sei.cmu.edu/library/organizational-models-for-computer-security-incident-response-teams-csirts/.

[1076] Enoch Agyepong, Yulia Cherdantseva, Philipp Reinecke, and Pete Burnap, "Towards a Framework for Measuring the Performance of a Security Operations Center Analyst," in *2020 International Conference on Cyber Security and Protection of Digital Services* (IEEE, 2020), https://doi.org/10.1109/CyberSecurity49315.2020.9138872.

[1077] Conrad Shayo and Frank Lin, "An Exploration of the Evolving Reporting Organizational Structure for the Chief Information Security Officer (CISO) Function," *Journal of Computer Science and Information*

Technology 7, no. 1 (2019), www.researchgate.net/publication/335404794_An_Exploration_of_the_Evolving_Reporting_Organizational_Structure_for_the_Chief_Information_Security_Officer_CISO_Function.

[1078] Daniel L. Costa, Michael J. Albrethsen, Matthew L. Collins, Samuel Perl, George Silowash, and Derrick Spooner, *An Insider Threat Indicator Ontology* (Software Engineering Institute, Carnegie Mellon University, May 2016), www.sei.cmu.edu/library/an-insider-threat-indicator-ontology/.

[1079] *Insider Threat Mitigation Guide* (CISA, November 2020), https://www.cisa.gov/resources-tools/resources/insider-threat-mitigation-guide.

[1080] "Ongoing Personnel Security: A Good Practice Guide," UK National Protective Security Authority, last modified December 4, 2023, www.npsa.gov.uk/specialised-guidance/insider-risk-guidance/ongoing-personnel-security.

[1081] Amruta Ambre and Narendra Shekokar, "Insider Threat Detection Using Log Analysis and Event Correlation," *Procedia Computer Science* 45 (2015): 436–445, https://doi.org/10.1016/j.procs.2015.03.175.

[1082] Jun Zengy, Xiang Wang, Jiahao Liu, Yinfang Chen, Zhenkai Liang, and Tat-Seng Chua, "SHADEWATCHER: Recommendation-Guided Cyber Threat Analysis Using System Audit Records," in *2022 IEEE Symposium on Security and Privacy* (IEEE, 2022), 489–506, https://doi.org/10.1109/SP46214.2022.9833669.

[1083] "User Access Reviews: A Step-by-Step Guide, Best Practices + Checklist," Drata, June 13, 2025, https://drata.com/blog/user-access-review.

[1084] "Network Security Logging and Monitoring — ITSAP.80.085," Canadian Centre for Cyber Security, Government of Canada, last modified December 22, 2022, www.cyber.gc.ca/en/guidance/network-security-logging-monitoring-itsap80085.

[1085] Gilberto Fernandes Jr., Joel J. P. C. Rodrigues, Luiz Fernando Carvalho, Jalal F. Al-Muhtadi, and Mario Lemes Proença Jr., "A Comprehensive Survey on Network Anomaly Detection," *Telecommunication Systems* 70 (2019): 447–489, https://doi.org/10.1007/s11235-018-0475-8.

[1086] Adrian Komadina, Mislav Martinić, Stjepan Groš, and Željka Mihajlović, "Comparing Threshold Selection Methods for Network Anomaly Detection," *IEEE Access* 12 (2024): 124943–124973, https://doi.org/10.1109/ACCESS.2024.3452168.

[1087] Arman A. Attar, Kaibin Bao, Veit Hagenmeyer, Tagir Fabarisov, and Andrey Morozov, "Improving Anomaly Detection with Adaptive Dynamic Threshold: A Review and Enhanced Method," in *2024 8th International Conference on System Reliability and Safety* (IEEE, 2024), 662–666, https://doi.org/10.1109/ICSRS63046.2024.10927575.

[1088] *Compromised Personal Network Indicators and Mitigations* (NSA, September 2020), https://media.defense.gov/2020/Sep/17/2002499615/-1/-1/0/COMPROMISED_PERSONAL_NETWORK_INDICATORS_AND_MITIGATIONS_20200914_FINAL.PDF.pdf.

[1089] Zi Long, Lianzhi Tan, Shenping Zhou, Chaoyang He, and Xin Liu, "Collecting Indicators of Compromise from Unstructured Text of Cybersecurity Articles Using Neural-Based Sequence Labelling," in *2019 International Joint Conference on Neural Networks* (IEEE, 2019), 1–8, https://doi.org/10.1109/IJCNN.2019.8852142.

[1090] Lisa Ehrlinger and Wolfram Wöß, "A Survey of Data Quality Measurement and Monitoring Tools," *Frontiers in Big Data* 5 (2022), https://doi.org/10.3389/fdata.2022.850611.

[1091] Nikhil Bangad, Vivekananda Jayaram, Manjunatha Sughaturu Krishnappa et al., "A Theoretical Framework for AI-Driven Data Quality Monitoring in High-Volume Data Environments," *International Journal of Computer Engineering and Technology* 15, no. 5 (2024): 618–636, https://doi.org/10.5281/zenodo.13878755.

[1092] William Fisher, R. Eugene Craft, Michael Ekstrom, Julian Sexton, and John Sweetnam, *Data Confidentiality: Detect, Respond to, and Recover from Data Breaches* (NIST, February 2024), https://doi.org/10.6028/NIST.SP.1800-29.

[1093] "MITRE Privacy Continuous Monitoring Framework," MITRE, October 8, 2019, www.mitre.org/sites/default/files/2021-11/pr-19-00598-6-privacy-continuous-monitoring-framework-briefing.pdf.

[1094] Georgios V. Lioudakis, Fotios Gogoulos, Anna Antonakopoulou, Dimitra I. Kaklamani, and Iakovos S. Venieris, "Privacy Protection in Passive Network Monitoring: An Access Control Approach," in *2009 International Conference on Advanced Information Networking and Applications Workshops* (IEEE, 2009), 109–116, https://doi.org/10.1109/WAINA.2009.158.

[1095] Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer, "BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices," *Advances in Neural Information Processing Systems* 37 (2024), https://proceedings.neurips.cc/paper_files/paper/2024/hash/26889e8359e7ef8a7f5d77457364ca55-Abstract-Datasets_and_Benchmarks_Track.html.

[1096] Jean Feng, Adarsh Subbaswamy, Alexej Gossmann et al., "Designing Monitoring Strategies for Deployed Machine Learning Algorithms: Navigating Performativity Through a Causal Lens," *Proceedings of Machine Learning Research* 236 (2024): 587–608, https://proceedings.mlr.press/v236/feng24a.html.

[1097] Fábio Pinto, Marco O. P. Sampaio, and Pedro Bizarro, "Automatic Model Monitoring for Data Streams," arXiv preprint arXiv:1908.04240 (2019), https://doi.org/10.48550/arXiv.1908.04240.

[1098] Olumuyiwa Ibidunmoye, Francisco Hernández-Rodriguez, and Erik Elmroth, "Performance Anomaly Detection and Bottleneck Identification," *ACM Computing Surveys* 48, no. 1 (September 2015), https://doi.org/10.1145/2791120.

[1099] Jose Morales, Luiz Antunes, Patrick Earl et al., "Insights on Implementing a Metrics Baseline for Post-deployment AI Container Monitoring," in *ICSSP '24: Proceedings of the 2024 International*

Conference on Software and Systems Processes (Association for Computing Machinery, 2024), 46–55, https://doi.org/10.1145/3666015.3666018.

[1100] Thomas A. Henzinger, Konstantin Kueffner, Vasu Singh, and I. Sun, "Alignment Monitoring," in *Runtime Verification* (Springer, 2025), 140–159, https://link.springer.com/chapter/10.1007/978-3-032-05435-7_9.

[1101] Nandita Bhaskhar, Daniel L. Rubin, and Christopher Lee-Messer, "An Explainable and Actionable Mistrust Scoring Framework for Model Monitoring," *IEEE Transactions on Artificial Intelligence* 5, no. 4 (April 2024): 1473–1485, https://doi.org/10.1109/TAI.2023.3272876.

[1102] Jiaming Ji, Wenqi Chen, Kaile Wang et al., "Mitigating Deceptive Alignment via Self-Monitoring," arXiv preprint arXiv:2505.18807 (2025), https://doi.org/10.48550/arXiv.2505.18807.

[1103] Avijit Ghosh, Aalok Shanbhag, and Christo Wilson, "FairCanary: Rapid Continuous Explainable Fairness," in *AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery, 2022), 307–316, https://doi.org/10.1145/3514094.3534157.

[1104] Thomas Henzinger, Mahyar Karimi, Konstantin Kueffner, and Kaushik Mallik, "Runtime Monitoring of Dynamic Fairness Properties," in *FAccT '23: 2023 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2023), 1–11, https://doi.org/10.1145/3593013.3594028.

[1105] Thomas A. Henzinger, Mahyar Karimi, Konstantin Kueffner and Kaushik Mallik, "Monitoring Algorithmic Fairness," in *Computer Aided Verification* (Springer, 2023), https://doi.org/10.1007/978-3-031-37703-7_17.

[1106] Yusheng Zheng, Yanpeng Hu, Tong Yu, and Andi Quinn, "AgentSight: System-Level Observability for AI Agents Using eBPF," in *PACMI '25: Proceedings of the 4th Workshop on Practical Adoption Challenges of ML for Systems* (Association for Computing Machinery, 2025), 110–115, https://doi.org/10.1145/3766882.3767169.

[1107] Bocheng Chen, Guangjing Wang, Hanqing Guo, Yuanda Wang, and Qiben Yan, "Understanding Multi-turn Toxic Behaviors in Open-Domain Chatbots," in *RAID '23: Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses* (Association for Computing Machinery, 2023), 282–296, https://doi.org/10.1145/3607199.3607237.

[1108] Mark Russinovich, Ahmed Salem, and Ronen Eldan, "Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack," arXiv preprint arXiv:2404.01833 (2024), https://doi.org/10.48550/arXiv.2404.01833.

[1109] Ahmed M. Fawaz and William H. Sanders, "Learning Process Behavioral Baselines for Anomaly Detection," in *2017 IEEE 22nd Pacific Rim International Symposium on Dependable Computing* (IEEE, 2017), 145–154, https://doi.org/10.1109/PRDC.2017.28.

[1110] Mathilde Machin, Jérémie Guiochet, Hélène Waeselynck, Jean-Paul Blanquart, Matthieu Roy, and Lola Masson, "SMOF: A Safety Monitoring Framework for Autonomous Systems," *IEEE Transactions*

on Systems, Man, and Cybernetics: Systems 48, no. 5 (May 2018): 702–715, https://doi.org/10.1109/TSMC.2016.2633291.

[1111] Xue Yang, Enda Howley, and Michael Schukat, "Agent-Based Dynamic Thresholding for Adaptive Anomaly Detection Using Reinforcement Learning," *Neural Computing and Applications* 37 (2025): 18775–18791, https://doi.org/10.1007/s00521-024-10536-0.

[1112] Younghun Chae, Natallia Katenka, and Lisa DiPippo, "An Adaptive Threshold Method for Anomaly-Based Intrusion Detection Systems," in *2019 IEEE 18th International Symposium on Network Computing and Applications* (IEEE, 2019), 1-4, https://doi.org/10.1109/NCA.2019.8935045.

[1113] Madhulika Srikumar, "Prioritizing Real-Time Failure Detection in AI Agents," Partnership on AI, September 11, 2025, https://partnershiponai.org/resource/prioritizing-real-time-failure-detection-in-ai-agents/.

[1114] Seyyed Ahmad Javadi, Chris Norval, Richard Cloete, and Jatinder Singh, "Monitoring AI Services for Misuse," in *AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery, 2021), 597–607, https://doi.org/10.1145/3461702.3462566.

[1115] *Cloud Security Playbook*, vol. 1 (U.S. Department of Defense, February 11, 2025), www.dmi-ida.org/knowledge-base-detail/Cloud-Security-Playbook-Volume-1.

[1116] Zara Abrams, "Using Generic AI Chatbots for Mental Health Support: A Dangerous Trend," American Psychological Association Services Inc., March 12, 2025, www.apaservices.org/practice/business/technology/artificial-intelligence-chatbots-therapists.

[1117] Maria Teresa Baldassarre, Danilo Caivano, Berenice Fernandez Nieto, Domenico Gigante, and Azzurra Ragone, "The Social Impact of Generative AI: An Analysis on ChatGPT," in *GoodIT '23: Proceedings of the 2023 ACM Conference on Information Technology for Social Good* (Association for Computing Machinery, 2023), 363–373, https://doi.org/10.1145/3582515.3609555.

[1118] Tamara Kneese and Emma Strubell, "A Holistic Framework for Measuring and Reporting AI's Impacts to Build Public Trust and Advance AI," Federation of American Scientists, June 26, 2025, https://fas.org/publication/reporting-ai-impact-to-build-public-trust/.

[1119] Yuelin Han, Zhifeng Wu, Pengfei Li, Adam Wierman, and Shaolei Ren, "The Unpaid Toll: Quantifying and Addressing the Public Health Impact of Data Centers," arXiv preprint arXiv:2412.06288 (2024), https://doi.org/10.48550/arXiv.2412.06288.

[1120] Sue Anne Teo, "Artificial Intelligence and Its 'Slow Violence' to Human Rights," *AI and Ethics* 5 (2025): 2265–2280, https://doi.org/10.1007/s43681-024-00547-x.

[1121] *AI for Impact: Strengthening AI Ecosystems for Social Innovation* (World Economic Forum, September 2024), www.weforum.org/publications/ai-for-impact-strengthening-ai-ecosystems-for-social-innovation/.

[1122] Aspen Hopkins, Isabella Struckman, Kevin Klyman, and Susan S. Silbey, "Recourse, Repair, Reparation, & Prevention: A Stakeholder Analysis of AI Supply Chains," in *FAccT '25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2025), 209–227, https://doi.org/10.1145/3715275.3732017.

[1123] Bureau of Cyberspace and Digital Policy, "Risk Management Profile for Artificial Intelligence and Human Rights," U.S. Department of State, July 25, 2024, https://2021-2025.state.gov/risk-management-profile-for-ai-and-human-rights/.

[1124] *Operational Guidance on Remediation and Access to Remedy* (Accountability Framework Initiative, May 2020), https://accountability-framework.org/the-accountability-framework/download-the-full-framework/.

[1125] Sijia Xiao, Haodi Zou, Alice Qian Zhang et al., "What Comes After Harm? Mapping Reparative Actions in AI Through Justice Frameworks," arXiv preprint arXiv:2506.05687 (2025), https://doi.org/10.48550/arXiv.2506.05687.

[1126] Chelsea Foushee Conard, "Quantifying the Severity of a Cybersecurity Incident for Incident Reporting" (graduate thesis, MIT, September 2024), https://hdl.handle.net/1721.1/157124.

[1127] "Harm Taxonomy," AI Incident Tracker, MIT AI Risk Initiative, accessed October 18, 2025, https://airisk.mit.edu/ai-incident-tracker/harm-taxonomy.

[1128] Chen Zhong, John Yen, Peng Liu, and Robert F. Erbacher, "Learning from Experts' Experience: Toward Automated Cyber Security Data Triage," *IEEE Systems Journal* 13, no. 1 (March 2019): 603–614, https://doi.org/10.1109/JSYST.2018.2828832.

[1129] Tao Ban, Takeshi Takahashi, Samuel Ndichu, and Daisuke Inoue, "Breaking Alert Fatigue: AI-Assisted SIEM Framework for Effective Incident Response," *Applied Sciences* 13, no. 11 (2023): 6610, https://doi.org/10.3390/app13116610.

[1130] Rick van der Kleij, Jan Maarten Schraagen, Beatrice Cadet, and Heather Young, "Developing Decision Support for Cybersecurity Threat and Incident Managers," *Computers and Security* 113 (2022), www.sciencedirect.com/science/article/pii/S016740482100359X.

[1131] Brittany Manley and David McIntire, *A Guide to Effective Incident Management Communications* (Software Engineering Institute, Carnegie Mellon University, February 2021), www.sei.cmu.edu/library/guide-to-effective-incident-management-communications/.

[1132] Marcos Osorno, Thomas Millar, and Danielle Rager, "Coordinated Cybersecurity Incident Handling: Roles, Processes, and Coordination Networks for Crosscutting Incidents," paper presented at the 16th International Command and Control Research and Technology Symposium, Quebec City, Canada, June 2011, https://apps.dtic.mil/sti/html/tr/ADA547075/.

[1133] "Insider Events: A Communications Guide to Reduce Their Impact," UK National Protective Security Authority, last modified March 4, 2024, www.npsa.gov.uk/specialised-guidance/insider-risk-guidance/insider-events-communications-guidance.

[1134] "Federal Incident Notification Guidelines," CISA, accessed September 7, 2025, www.cisa.gov/federal-incident-notification-guidelines.

[1135] Ren Bin Lee Dixon and Heather Frase, *An Argument for Hybrid AI Incident Reporting: Lessons Learned from Other Incident Reporting Systems* (CSET, March 2024), https://cset.georgetown.edu/publication/an-argument-for-hybrid-ai-incident-reporting/.

[1136] Sven Cattell, Avijit Ghosh, and Lucie-Aimée Kaffee, "Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities," *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (2024): 267–280, https://doi.org/10.1609/aies.v7i1.31635.

[1137] Richard Knight and Jason R. C. Nurse, "A Framework for Effective Corporate Communication After Cyber Security Incidents," *Computers & Security* 99 (2020), https://doi.org/10.1016/j.cose.2020.102036.

[1138] Violet Turri and Rachel Dzombak, "Why We Need to Know More: Exploring the State of AI Incident Documentation Practices," in *AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery, 2023), 576–583, https://doi.org/10.1145/3600211.3604700.

[1139] "Incident Response Team Depth Chart: Roles & Responsibilities," Wiz, August 23, 2024, www.wiz.io/academy/incident-response-team/.

[1140] Judith M. Brown, Steven Greenspan, and Robert Biddle, "Incident Response Teams in IT Operations Centers: The T-TOCs Model of Team Functionality," *Cognition, Technology, and Work* 18 (2016): 695–716, https://doi.org/10.1007/s10111-016-0374-2.

[1141] Lauren McIlvenny, "Leading AI Security Incident Response," Software Engineering Institute, Carnegie Mellon University, accessed September 2025, www.sei.cmu.edu/annual-reviews/2024-year-in-review/leading-ai-security-incident-response/.

[1142] Julie Steinke, Balca Bolunmez, Laura Fletcher, Vicki Wang, Alan J. Tomassetti, and Kristin M. Repchick, "Improving Cybersecurity Incident Response Team Effectiveness Using Teams-Based Research," *IEEE Security & Privacy* 13, no. 4 (2015): 20–29, https://doi.org/10.1109/MSP.2015.71.

[1143] Jonathan M. Spring and Phyllis Illari, "Review of Human Decision-Making During Computer Security Incident Analysis," *Digital Threats* 2, no. 2 (June 2021), https://doi.org/10.1145/3427787.

[1144] Subigya Nepal, Javier Hernandez, Robert Lewis et al., "Burnout in Cybersecurity Incident Responders: Exploring the Factors that Light the Fire," *Proceedings of the ACM on Human-Computer Interaction* 8, no. 1 (April 2024), https://doi.org/10.1145/3637304.

[1145] Rajesh Ganesan and Ankit Shah, "A Strategy for Effective Alert Analysis at a Cyber Security Operations Center," in *From Database to Cyber Security*, eds. Pierangela Samarati, Indrajit Ray, and Indrakshi Ray (Springer, 2018), https://doi.org/10.1007/978-3-030-04834-1_11.

[1146] "Incident Response Plan (IRP) Basics," CISA, January 31, 2024, www.cisa.gov/resources-tools/resources/incident-response-plan-irp-basics.

[1147] Keri K. Stephens, Ashley K. Barrett, and Michael J. Mahometa, "Organizational Communication in Emergencies: Using Multiple Channels and Sources to Combat Noise and Capture Attention," *Human Communication Research* 39, no. 2 (1 April 2013): 230–251, https://doi.org/10.1111/hcre.12002.

[1148] Neal Wagner, Cem Ş. Şahin, Michael Winterrose, James Riordan, Jaime Pena, and Diana Hanson, "Towards Automated Cyber Decision Support: A Case Study on Network Segmentation for Security," in *IEEE Symposium Series on Computational Intelligence* (IEEE, 2016), 1–10, https://doi.org/10.1109/SSCI.2016.7849908.

[1149] Jiehao Zhang, Simin Li, Weiwei Huang, Haoxin Jing, Qin Zhang, and Xing Xia, "Design and Computational Modeling of an AI-Based Automated Cybersecurity Incident Response System," *IEEE Access* 13 (2025): 154383–154394, https://doi.org/10.1109/ACCESS.2025.3603975.

[1150] Joe O'Brien, Shaun Ee, and Zoe Williams, *Deployment Corrections: An Incident Response Framework for Frontier AI Models* (Institute for AI Policy and Strategy, September 30, 2023), www.iaps.ai/research/deployment-corrections.

[1151] Eric C. Thompson, "Eradication, Recovery, and Post-incident Review," in *Cybersecurity Incident Response* (Apress, 2018), https://doi.org/10.1007/978-1-4842-3870-7_9.

[1152] *The Guide to Cyber Investigations*, 3rd ed. (Global Investigations Review, 2023), https://globalinvestigationsreview.com/guide/the-guide-cyber-investigations-archived/third-edition.

[1153] Abdul Rehman Javed, Waqas Ahmed, Mamoun Alazab, Zunera Jalil, Kashif Kifayat, and Thippa Reddy Gadekallu, "A Comprehensive Survey on Computer Forensics: State-of-the-Art, Tools, Techniques, Challenges, and Future Directions," *IEEE Access* 10 (2022): 11065–11089, https://doi.org/10.1109/ACCESS.2022.3142508.

[1154] Johannes Schneider and Frank Breitinger, "Towards AI Forensics: Did the Artificial Intelligence System Do It?," *Journal of Information Security and Applications* 76 (2023), https://doi.org/10.1016/j.jisa.2023.103517.

[1155] Carson Ezell, Xavier Roberts-Gaal, and Alan Chan, "Incident Analysis for AI Agents," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, no. 1 (2025): 865–878, https://doi.org/10.1609/aies.v8i1.36596.

[1156] Dipo Dunsin, Mohamed C. Ghanem, Karim Ouazzane, and Vassil Vassilev, "A Comprehensive Analysis of the Role of Artificial Intelligence and Machine Learning in Modern Digital Forensics and Incident Response," *Forensic Science International: Digital Investigation* 48 (2024), https://doi.org/10.1016/j.fsidi.2023.301675.

[1157] "Technical Approaches to Uncovering and Remediating Malicious Activity," CISA, last modified September 24, 2020, www.cisa.gov/news-events/cybersecurity-advisories/aa20-245a.

[1158] Quentin E. Hodgson, Aaron Clark-Ginsberg, Zachary Haldeman, Andrew Lauland, and Ian Mitch, *Managing Response to Significant Cyber Incidents: Comparing Event Life Cycles and Incident Response Across Cyber and Non-cyber Events* (RAND Corporation, May 2022), www.rand.org/pubs/research_reports/RRA1265-4.html.

[1159] "Data Breach Response: A Guide for Business," U.S. Federal Trade Commission, accessed September 2025, www.ftc.gov/business-guidance/resources/data-breach-response-guide-business.

[1160] "#StopRansomware Guide," Stop Ransomware, CISA, accessed October 2025, www.cisa.gov/stopransomware/ransomware-guide.

[1161] "Security Breach Notification Laws," National Conference of State Legislatures, last modified January 17, 2022, www.ncsl.org/technology-and-communication/security-breach-notification-laws.

[1162] *Cost of a Data Breach Report 2025: The AI Oversight Gap* (IBM, 2025), www.ibm.com/reports/data-breach.

[1163] "ISO 27001 Control 8.13: Information Backup," ISEO Blue, accessed October 2025, https://iseoblue.com/post/iso-27001-control-8-13-information-backup/.

[1164] Timothy McBridge, Michael Ekstrom, Lauren Lusty, Julian Sexton, and Anne Townsend, *Data Integrity: Recovering from Ransomware and Other Destructive Events* (NIST, September 2017), www.nccoe.nist.gov/publication/1800-11/VolB/index.html.

[1165] Sai Niveditha Varayogula, Kiranbhai Dodiya, Parth Lakhalani, and Arushi Chawla, "Computer Forensics Data Recovery Software: A Comparative Study," *International Journal of Innovative Research in Computer Science & Technology* 10, no. 2 (March 2022): 513–518, www.researchgate.net/publication/382411368_Computer_Forensics_Data_Recovery_Software_A_Comparative_Study.

[1166] Jonas Plum and Andreas Dewald, "Forensic APFS File Recovery," in *ARES '18: Proceedings of the 13th International Conference on Availability, Reliability and Security* (Association for Computing Machinery, 2018), https://doi.org/10.1145/3230833.3232808.

[1167] Ram Shankar Siva Kumar, David O Brien, Kendra Albert, Salomé Viljöen, and Jeffrey Snover, "Failure Modes in Machine Learning Systems," arXiv preprint arXiv:1911.11034 (2019), https://doi.org/10.48550/arXiv.1911.11034.

[1168] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova, "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores," in *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2020), 1–12, https://doi.org/10.1145/3313831.3376638.

[1169] Christopher Miller, Jay Shively, Summer Brandt, Helen Wauck, Vasanth Sarathy, and Richard Freedman, "Human as Automation Failsafe: Concept, Implications, Guidelines and Innovations," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 66, no. 1 (2022): 100–104, https://doi.org/10.1177/1071181322661014.

[1170] David Manheim and Aidan Homewood, "Limits of Safe AI Deployment: Differentiating Oversight and Control," arXiv preprint arXiv:2507.03525 (2025), https://doi.org/10.48550/arXiv.2507.03525.

[1171] Arighna Roy and Simone A. Ludwig, "Recovery Algorithm to Correct Silent Data Corruption of Synaptic Storage in Convolutional Neural Networks," *International Journal of Hybrid Intelligent Systems* 16, no. 3 (2020), https://doi.org/10.3233/HIS-200278.

[1172] Qi Liu and Wanjing Ma, "Navigating Data Corruption in Machine Learning: Balancing Quality, Quantity, and Imputation Strategies," *Future Internet* 17, no. 6 (2025): 241, https://doi.org/10.3390/fi17060241.

[1173] Baturalp Yalcin, Haixaing Zhang, Javad Lavaei, and Murat Arcak, "Exact Recovery for System Identification with More Corrupt Data than Clean Data," *IEEE Open Journal of Control Systems* 4 (2025): 1–17, https://doi.org/10.1109/OJCSYS.2024.3507452.

[1174] Avinash Maurya, Robert Underwood, M. Mustafa Rafique, Franck Cappello, and Bogdan Nicolae, "DataStates-LLM: Lazy Asynchronous Checkpointing for Large Language Models," arXiv preprint arXiv:2406.10707 (2024), https://arxiv.org/abs/2406.10707.

[1175] Richard Koo and Sam Toueg, "Checkpointing and Rollback-Recovery for Distributed Systems," *IEEE Transactions on Software Engineering* SE-13, no. 1 (1987): 23–31, https://doi.org/10.1109/TSE.1987.232562.

[1176] "Comprehensive Guide to IT Redundancy and Why It Is Essential," Giva, April 2, 2024, www.givainc.com/blog/it-redundancy/.

[1177] Patricia T. Endo, Moisés Rodrigues, Glauco E. Gonçalves, Judith Kelner, Djamel H. Sadok, and Calin Curescu, "High Availability in Clouds: Systematic Review and Research Challenges," *Journal of Cloud Computing* 5, no. 16 (2016), https://doi.org/10.1186/s13677-016-0066-8.

[1178] Elika Somani, Anjay Friedman, Henry Wu et al., *Strengthening Emergency Preparedness and Response for AI Loss of Control Incidents (*RAND, July 2025), www.rand.org/pubs/research_reports/RRA3847-1.html.

[1179] Mark D. Pogainy, "A Cyberattack and Its Aftermath: A Case Study of Survival," *American Bankruptcy Institute Journal*, September 2023, www.abi.org/abi-journal/a-cyberattack-and-its-aftermath-a-case-study-of-survival.

[1180] Daniel W. Pinkham, Ina M. Sala, Emilie T. Soisson, Brian Wang, and Matthew A. Deeley, "Are You Ready for a Cyberattack?," *Journal of Applied Clinical Medicine Physicians* 10, no. 22 (2021): 4–7, https://doi.org/10.1002/acm2.13422.

[1181] Dayse M. Cavalcanti, Publio M. M. Lima, Max H. de Queiroz, and Felipe G. Cabral, "Recovery of Discrete Event Systems After Active Cyberattacks," in *IEEE Control Systems Letters*, vol. 9 (IEEE, 2025), 1171–1176, https://doi.org/10.1109/LCSYS.2025.3580455.

[1182] Wei-Tsong Wang and S. Belardo, "Strategic Integration: A Knowledge Management Approach to Crisis Management," *Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (IEEE, 2005), 252a–252a, https://doi.org/10.1109/HICSS.2005.559.

[1183] John F. Preble, "Integrating the Crisis Management Perspective into the Strategic Management Process," *Journal of Management Studies* 34, no. 5 (2003), https://doi.org/10.1111/1467-6486.00071.

[1184] Epaminondas Koronis and Stavros Ponis, "A Strategic Approach to Crisis Management and Organizational Resilience," *Journal of Business Strategy* 39, no. 1 (2018): 32–42, https://doi.org/10.1108/JBS-10-2016-0124.

[1185] Merce Bernardo, Marti Casadesus, Stanislav Karapetrovic, and Iñaki Heras, "An Empirical Study on the Integration of Management System Audits," *Journal of Cleaner Production* 18, no. 5 (2010): 486–495, https://doi.org/10.1016/j.jclepro.2009.12.001.

[1186] Rob J. B. Vanwersch, Khurram Shahzad, Irene Vanderfeesten et al., "A Critical Evaluation and Framework of Business Process Improvement Methods," *Business & Information Systems Engineering* 58 (2016): 43–53, https://doi.org/10.1007/s12599-015-0417-x.

[1187] Atif Ahmad, Justin Hadgkiss, and A. B. Ruighaver, "Incident Response Teams—Challenges in Supporting the Organisational Security Function," *Computers & Security* 31, no. 5 (2012): 643–652, https://doi.org/10.1016/j.cose.2012.04.001.

[1188] Bob Violino, "How to Conduct an Effective Post-incident Review," CSO, June 20, 2025, www.csoonline.com/article/4009438/how-to-conduct-an-effective-post-incident-review.html.

[1189] Ghina Fitriya, Boy Sandi Kritian Sihombing, Fatoumatta Binta Jallow, Sofian Lusa, Nadya Safitri, and Dana Indra Sensuse, "Enhancing Post-incident Activities Through Knowledge Management Models: A Systematic Literature Review," *Indonesian Journal of Computer Science* 13, no. 6 (2024), https://doi.org/10.33022/ijcs.v13i6.4527.

[1190] Clare M. Patterson, Jason R. C. Nurse, and Virginia N. L. Franqueira, "Learning from Cyber Security Incidents: A Systematic Review and Future Research Agenda," *Computers & Security* 132 (2023), https://doi.org/10.1016/j.cose.2023.103309.

[1191] Dharani Goli, Hamad Al-Mohannadi, and Mohammad Shah, "Plan, Prepare and Respond: A Holistic Cyber Security Risk Management Platform," in *10th International Conference on Future Internet of Things and Cloud* (IEEE, 2023), 367–374, https://doi.org/10.1109/FiCloud58648.2023.00060.

[1192] Rajat Bhagwat and Milind Kumar Sharma, "Performance Measurement of Supply Chain Management: A Balanced Scorecard Approach," *Computers & Industrial Engineering* 53, no. 1 (2007): 43–62, https://doi.org/10.1016/j.cie.2007.04.001.

[1193] Peter Gilmour, "A Strategic Audit Framework to Improve Supply Chain Performance," *Journal of Business & Industrial Marketing* 14, no. 5–6 (1999): 355–366, https://doi.org/10.1108/08858629910290102.

[1194] Ceren Atilgan and Peter McCullen, "Improving Supply Chain Performance Through Auditing: A Change Management Perspective," *Supply Chain Management: An International Journal* 16, no. 1 (2011): 11–19, https://doi.org/10.1108/13598541111103467.

[1195] Ipek Deveci Kocakoc, and Ali Şen, "Utilising Surveys for Finding Improvement Areas for Customer Satisfaction Along the Supply Chain," *International Journal of Market Research* 48, no. 5 (2006): 623–636, https://doi.org/10.1177/147078530604800509.

[1196] Deidra J. Schleicher, Heidi M. Baumann, David W. Sullivan, and J. Yim, "Evaluating the Effectiveness of Performance Management: A 30-Year Integrative Conceptual Review," *Journal of Applied Psychology* 104, no. 7 (2019): 851–887, https://doi.org/10.1037/apl0000368.

[1197] Donald L. Kirkpatrick and James D. Kirkpatrick, *Evaluating Training Programs: The Four Levels*, 3rd ed. (Berrett-Koehler Publishers, 2006), https://books.google.com/books/about/Evaluating_Training_Programs.html?id=BJ4QCmvP5rcC.

[1198] Sunil Chaudhary, Vasileios Gkioulos, Sokratis Katsikas, "Developing Metrics to Assess the Effectiveness of Cybersecurity Awareness Program," *Journal of Cybersecurity* 8, no. 1 (2022), https://doi.org/10.1093/cybsec/tyac006.

[1199] Martin Tessmer, Daniel McCann, and Michael Ludvigsen, "Reassessing Training Programs: A Model for Identifying Training Excesses and Deficiencies," *Educational Technology Research and Development* 47 (1999): 86–99, https://doi.org/10.1007/BF02299468.

[1200] Jacopo Soldani and Antonio Brogi, "Anomaly Detection and Failure Root Cause Analysis in (Micro) Service-Based Cloud Applications: A Survey," *ACM Computing Surveys* 55, no. 3 (March 2023), https://doi.org/10.1145/3501297.

[1201] Marc Solé, Victor Muntés-Mulero, Annie Ibrahim Rana, and Giovani Estrada, "Survey on Models and Techniques for Root-Cause Analysis," arXiv preprint arXiv:1701.08546 (2017), https://doi.org/10.48550/arXiv.1701.08546.

[1202] Ihab Mohamed, Hesham A. Hefny, and Nagy R. Darwish, "Enhancing Cybersecurity Defenses: A Multicriteria Decision-Making Approach to MITRE ATT&CK Mitigation Strategy," arXiv preprint arXiv:2407.19222 (2024), https://doi.org/10.48550/arXiv.2407.19222.

[1203] Vincent C. Hu and Karen Scarfone, *Guidelines for Access Control System Evaluation Metrics* (NIST, September 2012), http://dx.doi.org/10.6028/NIST.IR.7874.

[1204] *Auditing Cybersecurity Operations: Prevention and Detection*, 2nd ed. (The Institute of Internal Auditors, 2025), www.theiia.org/en/content/guidance/recommended/supplemental/gtags/gtag-auditing-cybersecurity-operations-prevention-and-detection/.

[1205] "Cybersecurity Program Audit Guide," U.S. Government Accountability Office, September 28, 2023, www.gao.gov/products/gao-23-104705/.

[1206] "Data Audit vs. Data Impact Assessments: Understanding the Differences," GDPR Advisor, accessed October 25, 2025, www.gdpr-advisor.com/data-audit-vs-data-impact-assessment-understanding-the-differences/.

[1207] Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, and Tomasz Janowski, "Data Governance: Organizing Data for Trustworthy Artificial Intelligence," *Government Information Quarterly* 37, no. 3 (2020), https://doi.org/10.1016/j.giq.2020.101493.

[1208] Paolo Passeri, "The Risk of Accidental Data Exposure by Generative AI Is Growing," *Infosecurity Magazine*, August 16, 2023, www.infosecurity-magazine.com/blogs/accidental-data-exposure-gen-ai/.

[1209] Mike Van Stone, "Mistakes Happen—Mitigating Unintentional Data Loss," *ISACA Journal* 1 (January 17, 2018), www.isaca.org/resources/isaca-journal/issues/2018/volume-1/mistakes-happenmitigating-unintentional-data-loss.

[1210] Malek Mechergui and Sarath Sreedharan, "Goal Alignment: Re-analyzing Value Alignment Problems Using Human-Aware AI," *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 9 (2024): 10110–10118, https://doi.org/10.1609/aaai.v38i9.28875.

[1211] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh, "Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems," in *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2021), 598–609, https://doi.org/10.1145/3442188.3445921.

[1212] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, and Eric Horvitz, "Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance," *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, no. 1 (2019): 2–11, https://doi.org/10.1609/hcomp.v7i1.5285.

[1213] John McDermid, Kester Clegg, Yan Jia, and Ibrahim Habli, *AI Guardrails: Concepts, Models, and Methods* (Centre for Assuring Autonomy, University of York, 2024), www.york.ac.uk/media/assuring-autonomy/news/AI%20Guardrails%20CfAA%20White%20Paper.pdf.

[1214] Laura Freeman, "Test and Evaluation for Artificial Intelligence," *INSIGHT* 23, no. 1 (2020): 27–30, https://doi.org/10.1002/inst.12281.

[1215] Jaganmohan Chandrasekaran, Tyler Cody, Nicola McCarthy, Erin Lanus, Laura Freeman, and Kristen Alexander, "Testing Machine Learning: Best Practices for the Life Cycle," *Naval Engineers Journal* 136, no. 1–2 (2024): 249–263, www.ingentaconnect.com/content/asne/nej/2024/00000136/f0020001/art00039.

[1216] Tim A. Majchrzak, *Improving Software Testing: Technical and Organizational Developments* (Springer, 2012), https://doi.org/10.1007/978-3-642-27464-0.

[1217] Matt Pope and Jonathan Sillito, "Post-incident Action Items: Crossroads of Requirements Engineering and Software Evolution," *MO2RE 2024: Proceedings of the 1st IEEE/ACM Workshop on*

Multi-disciplinary, Open, and RElevant Requirements Engineering (Association for Computing Machinery, 2024), 1–7. https://doi.org/10.1145/3643666.3648578.

[1218] Bonnie Collier, Tom DeMarco, and Peter Fearey, "A Defined Process for Project Post Mortem Review," *IEEE Software* 13, no. 4, (July 1999): 65–72, https://doi.org/10.1109/52.526833.

[1219] Brian Fitzgerald and Klaas-Jan Stol, "Continuous Software Engineering and Beyond: Trends and Challenges," in *RCoSE 2014: Proceedings of the 1st International Workshop on Rapid Continuous Software Engineering* (Association for Computing Machinery, 2014), 1–9, https://doi.org/10.1145/2593812.2593813.

[1220] Katharina Weitz, Ruben Schlagowski, Elisabeth André, Maris Männiste, and Ceenu George, "Explaining It Your Way—Findings from a Co-creative Design Workshop on Designing XAI Applications with AI End-Users from the Public Sector," in *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2024), https://doi.org/10.1145/3613904.3642563.

[1221] Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzaeg, "The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users," in *CUI '23: Proceedings of the 5th International Conference on Conversational User Interfaces* (Association for Computing Machinery, 2023), https://doi.org/10.1145/3571884.3597144.

[1222] Asbjørn Følstad, Effie Law, and Kasper Hornbæk, "Analysis in Practical Usability Evaluation: A Survey Study," in *CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2012), 2127–2136, https://doi.org/10.1145/2207676.2208365.

[1223] Javier A. Bargas-Avila and Kasper Hornbæk, "Old Wine in New Bottles or Novel Challenges: A Critical Analysis of Empirical Studies of User Experience," in *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2011), 2689–2698, https://doi.org/10.1145/1978942.1979336.

[1224] Lisa Brand, Bernhard G. Humm, Andrea Krajewski, and Alexander Zender, "Towards Improved User Experience for Artificial Intelligence Systems," *Engineering Applications of Neural Networks, Communications in Computer and Information Science* 1826 (2023), https://doi.org/10.1007/978-3-031-34204-2_4.

[1225] World Economic Forum, "Towards Equitable AI: New Report Charts Path to AI Competitiveness," news release, January 21, 2025, www.weforum.org/press/2025/01/towards-equitable-ai-new-report-charts-path-to-ai-competitiveness/.