Issue Brief

Harmonizing AI Guidance

Distilling Voluntary Standards and Best Practices into a Unified Framework

Authors
Kyle Crichton
Abhiram Reddy
Jessica Ji
Ali Crawford
Mia Hoffmann
Colin Shea-Blymyer
John Bansemer

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

September 2025

Executive Summary

Organizations looking to develop or deploy artificial intelligence (AI) systems face many barriers in trying to operationalize AI-related best practices. In addition to the many practical hurdles to implementation, such as a lack of resources or in-house expertise, the complex landscape of existing AI guidance itself presents a substantial challenge. Organizations face an overload of information, coming from a myriad of disparate sources, that is often written in language that can be inaccessible to many organizations. This places an enormous burden on practitioners to sort through and decipher this guidance on their own—requiring time, resources, and expertise that many organizations, particularly smaller ones, cannot afford.

To address these challenges, the researchers at CSET have attempted to do this intensive work for organizations. In this report, we present a harmonized framework for how an organization should govern, manage, and protect its technology—and how to integrate emerging technologies such as artificial intelligence into its existing practices. This work distills more than 7,000 recommendations, collected from 52 different guidance documents, into a condensed set of 258 recommendations. These recommendations are grouped into 5 overarching categories and 34 topic areas, enabling organizations to quickly identify the most important practices across a broad set of disciplines. The breadth of content covered in this framework exceeds that of any existing individual guidance document. To match this scope, organizations would otherwise need more than 900 recommendations from seven or more different frameworks to approximate. In creating this framework, we develop a novel process for harmonization and methods to validate the results that can be reused for other applications.

Alongside each recommendation, we indicate the degree to which the content is developed from AI-specific guidance. This information, derived from the harmonization process, helps to illustrate how new AI guidance overlays with existing cybersecurity, privacy, and risk management practices. We conclude our analysis by identifying where current AI reports have been focused and highlighting gaps in existing knowledge, work that forthcoming CSET research aims to address.

Table of Contents

Introduction

With the rapid adoption of artificial intelligence, a myriad of reports have been produced in the last few years providing guidance on ensuring the safe, secure, and trustworthy use of AI systems. Organizations looking to implement these practices face the challenge of deciphering implementation guidance from a patchwork set of government-developed frameworks, technical research reports, and disparate industry practices. Faced with a breadth of information, practitioners must piece together how these various recommendations fit together, not only with each other but also existing cybersecurity, privacy, and risk management practices. These demands place a substantial burden on organizations—requiring time, resources, and expertise that many organizations, particularly smaller ones, cannot afford. As a result, the inaccessibility of guidance may preclude many organizations from adopting AI technologies and risks the uneven implementation of safety and security measures across the organizations that do.

To address these barriers, this report deciphers and harmonizes existing guidance so that organizations do not have to undertake these efforts. Our work provides practitioners with a single, clearly written, and streamlined set of recommendations that covers the scope of safety, security, privacy, and risk management practices. In consolidating recommendations from existing guidance across these disciplines, we provide a more manageable set of practices for organizations to implement and we identify the areas that organizations should prioritize when developing or deploying AI systems. This harmonized framework represents the first of three stages of research—which will be presented in a series of CSET reports—that seeks to 1) harmonize, 2) operationalize, and 3) tailor best practices to facilitate the adoption of safe, secure, and trustworthy AI.

To develop this framework, we synthesized a harmonized set of recommendations based on an analysis of 52 existing guidance documents. Of these reports, 29 provide recommendations for organizations developing or adopting AI systems. In addition, we include 23 non-AI reports that cover a range of closely related topics to better understand how AI guidance aligns with, and can be integrated into, existing organizational practices. Collectively, these reports were developed by a range of international bodies, government agencies both in the United States and abroad, standards-setting organizations, academic institutions, industry associations, and private companies. We applied a mixture of quantitative and qualitative methods to distill the 7,741 recommendations extracted from these reports into 258 that capture the most salient information while retaining the breadth of topics covered to the extent

it was feasible. The framework covers 34 topics areas, organized into five high-level groups: Governance, Safety, Security, Privacy, and Detection & Response. To ensure that the framework can be a useful standalone resource for organizations we provide evidence to validate the accuracy, representativeness, and completeness of the consolidated set of recommendations.

Alongside each recommendation in our framework, we indicate the degree to which the content is based on guidance from AI-specific reports. This information, derived as a part of the harmonization process developed by CSET researchers, illustrates how new AI guidance overlays with existing organizational practices. Building on these results, we also provide an assessment of where AI guidance, to date, has been focused. We find that these recommendations center on the broader set of risks and impacts stemming from AI systems, a greater need for transparency, novel security vulnerabilities, more extensive testing and evaluation, and new ethical considerations related to synthetic content. Finally, we identify several gaps in existing AI guidance where further work is needed.

In this report, we start by providing background information on the state of AI guidance and challenges organizations face in implementation. We then detail our methodology, provide the results of our analysis, and present the final harmonized framework. We close by discussing the implications of existing guidance for the implementation of AI safety and security practices. Practitioners seeking practical guidance should feel free to skip directly to the **harmonized framework** which starts on page 25.

Background

Calls for developing standards and best practices for AI and machine learning (ML)-based systems long predates the AI boom following the popularization of transformer-based models such as ChatGPT in 2022. Prior to 2020, more than 80 organizations had published some form of AI principles or ethical guidelines.[1] Following the explosion of interest in AI, renewed calls for developing practical standards and best practices have grown. Chief among that crowded set of voices, former President Joe Biden's Executive Order on Artificial Intelligence called for the development and use of "safe, secure, and trustworthy" AI.[2] Since its release, a plethora of AI frameworks, best practices, and reports have been published within the United States and internationally. These guidance documents have been developed by a wide range of organizations—including international bodies, government agencies, standards-setting organizations, academic institutions, industry associations, and private companies—and cover a broad spectrum of topics related to AI. Despite these persistent efforts, there remains a substantial gap in determining how to implement these recommendations.[3]

A Divided and Shifting Landscape

It was no coincidence that the Biden administration's executive order on AI explicitly used the terms "safe, secure, and trustworthy" to lay out its goals for AI development and use, as the AI landscape has long been fractured along these lines. These complementary, yet at times seemingly competitive, disciplines have largely grown out of separate academic fields, each with their own communities and publishing venues. Issues pertaining to AI safety, such as alignment and robustness, have been the focus of machine learning researchers publishing at venues such as the Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence and the Conference on Neural Information Processing Systems (NeurIPS). AI security has largely fallen into the domain of cybersecurity and privacy researchers whose venues include the Institute of Electrical and Electronics Engineers (IEEE) Symposium on Security and Privacy and the USENIX Security Symposium. Finally, AI trustworthiness—a more nebulous term—encompasses a wide range of work related to bias and fairness, human rights and societal impacts, and the political economy of AI. This research can often be found at venues such as the Association for Computing Machinery (ACM) Conference on Fairness, Accountability, and Transparency (FAccT) and the AAAI/ACM Conference on AI, Ethics, and Society (AIES).

The differences in perspectives among these communities has been reflected in the ongoing debate over the scoping and mission of the U.S. and U.K. AI safety institutes,

both of which continue to evolve, with the former being rebranded to the Center for AI Standards and Innovation and the latter to the AI Security Institute.[4] Recent work has called the inclusion of both safety and security perspectives necessary for an organization's approach to AI risk management.[5] That should be taken a step further to include trust-related perspectives as well.

Today, the AI landscape continues to evolve and the second Trump administration adds further ambiguity to the situation. While the new administration's planned AI direction has not been fully expressed, it appears that it will, at least in part, be a departure from the previous administration's. Vice President JD Vance's speech at the Artificial Intelligence Action Summit held in Paris in February 2025 made clear that the administration is no longer focused on AI safety and indicated that federal AI regulation was likely off the table.[6] As instructed by President Trump in Executive Order 14179, "Removing Barriers to American Leadership in Artificial Intelligence," the Office of Management and Budget (OMB) revised their guidance on AI to align with new administrative priorities.[7] However, the new guidance released in April, OMB memoranda M-25-21 and M-25-22, takes a remarkably similar approach to the Biden administration's AI governance, OMB memoranda M-24-10 and M-24-18.[8] This suggests that the new administration's approach may be less of a divergence than recent rhetoric might indicate. Furthermore, there remains a strong desire among the private sector and government agencies for guidance on adopting and implementing AI systems. This combination of factors makes the development of voluntary standards and best practices potentially of even greater importance.

Challenges in Operationalizing AI Guidance

Organizations that are looking to develop or deploy AI systems face many barriers in trying to operationalize AI-related best practices. While there are many practical hurdles to implementation, such as limitations in the resources or AI-related expertise within the organization, the complex landscape of existing AI guidance presents a substantial impediment in and of itself. While not exhaustive, we provide an overview of several major challenges organizations face when trying to operationalize AI guidance:

- **Information Overload:** One of the core challenges in operationalizing AI guidance is navigating the myriad of different frameworks, best practices, and reports on AI produced just in the last two years. In aggregate, these documents present an enormous amount of information for organizations to ingest. Trying to keep up with the large and ever-growing scope of AI-related guidance is a tall task, much less trying to internalize and operationalize it. Distilling this

information into actionable steps requires time and expertise that many organizations, particularly smaller ones, do not have.

- **Disparate Sources:** Compounding the problem of information overload, existing AI-related guidance is spread across numerous overlapping, yet disparate documents. Organizations seeking to take a comprehensive approach to AI—addressing issues related to safety, security, risk management, privacy, and due diligence just to name a few—must gather relevant information from a variety of sources, both AI-specific and more general. Just using a subset of the guidance produced by National Institute of Standards and Technology (NIST) as an example, organizations must integrate recommendations from the AI Risk Management Framework (AI RMF), Cybersecurity Framework (CSF), and Privacy Framework into a cohesive enterprise strategy.[9] In the absence of overarching guidance, organizations are left on their own to figure out how new AI guidance fits together with existing organizational practices related to the management and oversight of technological resources. The lack of clear guidance on how to integrate these practices risks the development of uneven and patchwork implementations across organizations.

- **Inaccessible Language:** The recommendations provided in guidance documents are often challenging to understand. In general, these reports often contain complex syntax, vague terminology, technical jargon, and at times inscrutable language. For example, Manage 2.2 of the NIST AI RMF recommends that "mechanisms are in place and applied to sustain the value of deployed AI systems." While free from jargon, it is not readily apparent what "sustaining the value" of an AI system means. Another example, taken from NIST's Security and Privacy Controls for Information Systems (SP 800-53), reads: "Require personnel to associate and maintain the association of [Assignment: organization-defined security and privacy attributes] with [Assignment: organization-defined subjects and objects] in accordance with [Assignment: organization-defined security and privacy policies]." The complexity of certain recommendations and the implicit assumptions of certain background knowledge can make guidance documents less accessible to organizations, particularly for nontechnical audiences who may be responsible for managing the implementation of these recommendations.

- **Lack of Implementation Details:** The recommendations provided in guidance documents tend to be very high level, often providing a goal without details on how to achieve it. For example, ISO 23894, Artificial Intelligence Guidance on

Risk Management, recommends, "When identifying risks of AI systems, various risks sources should be taken into account depending on the nature of the system under consideration and its application context." While a range of risks are described in other sections of the report, the guidance does not elaborate on how an organization should identify relevant AI risks or account for them. That said, there are several logical reasons why guidance documents tend to refrain from providing implementation details. First, technology, particularly when it comes to AI, and related best practices change over time. Second, in the case of AI specifically, there remains substantial uncertainty about what best practices should be. Third, AI—applied broadly as the term is today—covers a wide range of ML-based applications and use cases. As a result, reports that offer granular implementation details run the risk of having to be updated more frequently as best practices change and uncertainties are resolved. This may contribute to reports being exceedingly long to account for a larger set of use cases (potentially exacerbating the problem of information overload). Yet, shying away from these details does a disservice to organizations that require this level of information to make practical use of guidance.

- **One-Size-Fits-All:** A secondary effect of providing only high-level guidance is that, in attempting to provide recommendations that apply to all use cases, they end up too broad to be useful. Colloquially referred to as the one-size-fits-all problem, guidance that is broadly applicable inherently fails to address the nuances in different use cases that are often critical in developing effective protections and practices. The variations can arise across different sectors, sizes and resources of organizations, applications in which AI is used, and types of AI models employed. Once again, it is up to organizations to determine how to apply general guidance to their specific applications.

If we want organizations to adopt AI technologies in a safe, secure, and trustworthy manner, we must lower the barriers to do so. The issues enumerated above are inherently in tension with one another. Providing granular implementation details adds to the problem of information overload. Similarly, tailoring guidance to different use cases across multiple reports contributes to an even wider breadth of sources. As a result, there is no perfect solution.

However, there are ways to make improvements, and existing efforts can help chart a way forward. Our approach, which we describe in detail in the next section, draws on lessons from ongoing work being done at NIST, the U.S. Department of Defense (DOD)'s Chief Digital and Artificial Intelligence Office (CDAO), and the Partnership on

AI (PAI). The first lesson comes from NIST's development of the AI RMF playbook as a means to provide greater implementation detail to the high-level recommendations provided in the AI RMF.[10] While we believe the playbook could be further improved by going into greater detail, the two documents provide a model for how broad and detailed guidance can be provided in tandem. The second lesson comes from NIST's use of framework profiles to tailor broad guidance to specific sectors or use cases.[11] This work is effort intensive but extremely valuable to organizations. The third lesson comes from the interactive interfaces provided by CDAO and PAI that enable users to customize and generate tailored guidance based on factors such as the type of AI models being used or the role of the user within the organization.[12] Beyond the advantage in customization, these toolkits also provide guidance in a much more accessible manner than static reports. These lessons serve as the foundation for our approach to operationalizing AI-related guidance.

Methods

This report is the first in a series of work produced by CSET that aims to address the challenges that existing AI guidance presents to organizations. We break this work down into three phases:

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

This report covers the first of these three stages: harmonization. We present a harmonized framework that distills the enormous amount of information contained within existing guidance documents into a much smaller, more manageable set of clearly written recommendations. At the same time, the framework retains coverage over the breadth of topics that organizations would previously have to source from multiple reports, integrating AI-specific guidance with that stemming from other disciplines, thereby providing a standalone resource for organizations. This framework helps to address the problems of information overload, inaccessible language, and disparate sources of information. To develop the harmonized set of recommendations, we first collate guidance from a range of existing reports and then apply a mixed quantitative and qualitative approach to synthesize the recommendations and validate the results. We describe this process in detail in the following sections.

Although our harmonized framework addresses several of the aforementioned challenges in adopting AI guidance, the recommendations we provide in this report remain high level and broadly applicable. In future CSET reports, we will build on this framework to tackle operationalizing and tailoring. In the second (operationalizing) phase, we will provide granular steps for organizations to implement the recommendations in this report. This work will draw on a broad review of academic research and industry best practices to identify a concrete set of actions, techniques, and tools that organizations can operationalize. In the vein of the NIST AI RMF Playbook, this aims to address the lack of implementation details in most guidance documents. In the third (tailoring) phase, we will apply our framework and the implementation details from the second phase to several different AI deployment cases, identifying which practices are most relevant and how they should be customized to meet application-specific needs. This work aims to mitigate the one-size-fits-all problem present in most guidance documents. This work will serve a similar role to NIST's profiles and we aim to provide this information in an accessible format like that used by CDAO and PAI.

In the following sections we describe in greater detail the methodology we used in collating and harmonizing existing guidance.

Collating Existing Guidance

We based our proposed framework on an analysis of 7,741 recommendations collected from 52 different guidance documents developed in the United States, the United Kingdom, the European Union, Japan, and Singapore. The reports were produced by a range of international bodies, government agencies, standards-setting organizations, industry associations, think tanks, and academic institutions. A full list of the guidance documents examined in our analysis can be found in the appendix.

To identify relevant guidance documents, we started with NIST's AI Risk Management Framework, Cybersecurity Framework, and Privacy Framework—three guidance documents that participants in a CSET workshop held in June 2024 identified as the foundation for their organizations' approach to AI.[13] Based on these reports, the scope of our analysis includes guidance related to AI, risk management, cybersecurity, and privacy. We then examined the publications of known actors in the AI space, identified common references from existing research reports, and performed a broad scan of the overall literature. We included guidance documents in our corpus if they met the following criteria:

1. The content of the report relates to one of the four identified topic areas (AI, risk management, cybersecurity, or privacy).
2. The document has an English-language version.
3. The content of the report is prescriptive rather than descriptive (i.e., the report provides recommendations, not just information).
4. The report is structured such that recommendations can be extracted individually. Frameworks that have hierarchical or bulleted recommendations are most conducive to this process.
5. The report is prominent in the shaping of organizational practices and industry standards, is highly referenced, or is produced by a well-established organization in the field.

Based on this scoping, the corpus of reports can be generally divided into two groups. The first group, consisting of 29 reports, provides recommendations tailored to organizations developing or adopting AI systems. This guidance represents a rough approximation of the collective knowledge of AI best practices developed to date. The second group, consisting of 23 reports, provides guidance that does not specifically relate to AI but covers cybersecurity, privacy, and risk management practices more

broadly. These reports represent a baseline of existing organizational practices related to the management and oversight of technology. We included these reports to better understand how AI-related guidance aligns with, and can be integrated into, these existing practices. Neither group of reports is exhaustive. In particular, the breadth of non-AI guidance it is quite expansive.

Harmonizing Recommendations

To harmonize recommendations across guidance documents, we used quantitative methods to cluster recommendations into groups based on similar topics and then used qualitative methods to extract the most salient recommendations within each group. This process enabled us to analyze the corpus in an objective, structured manner while also leveraging the expertise and knowledge of our team of researchers. Our methodology followed six steps:

1. **Extracting**: We extracted individual recommendations from each report. In most cases, individual recommendations correspond one-to-one with those in the source document. In some cases, lengthy recommendations were separated into multiple recommendations. For reports with hierarchical structures, we extracted individual recommendations at the most granular level possible. For example, with the AI RMF, recommendations correspond to the subcategory level (e.g., Govern 1.1) rather than the function (Govern) or category (Govern 1) level.

2. **Standardizing**: We standardized certain terminology, the use of references, and the grammatical voice used in the recommendations, as they vary widely across reports and we found this to have some effect on the initial clustering results. In particular, the use of the terms “artificial intelligence” or “machine learning” resulted in separate clusters whenever they appear. Because part of our goal in harmonization is to identify commonalities across AI and non-AI related guidance, we masked these terms with more generic ones, replacing them with either “system” or “technology,” depending on grammatical use. We replaced the terminology used to refer to different audiences (e.g., companies, government agencies, member states, etc.) with “organization.” We also removed all in-text references to external reports, other sections of the documents, and placeholder text. Finally, we converted all of the recommendations to the active, rather than passive, voice. An illustrative example of the standardization process is included in the appendix.

3. **Embedding**: We generated a vector embedding for each recommendation using application programming interface (API) calls to OpenAI's text-embedding-3-large model.[14] Each embedding provides a numeric representation (a vector of 3,072 high-precision decimal values) that corresponds to the text of the original recommendation. Recommendations related to similar topics result in embeddings that are closer in mathematical distance to one another—a principle underpinning the foundation of LLMs.[15]

4. **Clustering**: Using these vector embeddings, we grouped similar sets of recommendations using agglomerative clustering. We incrementally increased the number of clusters until new, logically grouped clusters stopped emerging from the dataset. We found that this occurred after 34 clusters are reached. Because agglomerative clustering is hierarchical, we used the results to group the 34 low-level clusters we refer to as "topics" into five higher-level groupings we refer to as "categories" for easier organization. For example, the Network Security and Physical Security topic clusters fall under the Security category. We used this structure to define the two-tier hierarchy of our framework. The visualizations presented in the results section use T-distributed stochastic neighbor embedding (t-SNE) to project the clustering results into a two-dimensional space.

5. **Qualitative Coding**: To develop the final set of recommendations, we employed qualitative coding methods to identify core concepts and commonalities within each of the 34 topic clusters. For each cluster, we iteratively developed a codebook—a set of descriptive labels that capture underlying themes and patterns in a dataset. We created these codebooks using emergent methods, a flexible technique that enables researchers to organize and structure data qualitatively when no preset codebook exists.[16] We then assigned one or more codes from the corresponding codebook to each recommendation. To validate the results, a separate member of the research team coded a subset of those recommendations. We found that this process results in fairly high intercoder reliability using Fuzzy Kappa (μ=0.616, σ=0.096), a version of Cohen's Kappa that allows for multiple codes per item, as our metric.[17] An average score of 0.616 indicates a substantial degree of agreement among coders.[18] We then used thematic analysis, where applicable, to reduce the number of codes to between 5 and 10 themes. For each theme, we synthesized a harmonized recommendation based on the set of recommendations associated with that theme. In this way, we captured the most salient recommendations and maintained a representative breadth across each topic cluster.

6. **Validating:** In addition to using Cohen's Kappa, we also empirically validated the accuracy, representativeness, and completeness of our results. Using the same process as described in step 3, we embedded the harmonized recommendations and projected the results into the same embedding space as the clustering results. We then visually presented evidence that supports the validity of the recommendations produced through this mixed-methods harmonization process.

Limitations

While we examined a broad range of reports in our analysis, our scope is by no means exhaustive. Our work focuses on voluntary best practices and excludes regulation and legislation. Therefore, following the guidance presented in the harmonized framework does not ensure compliance with existing legal requirements. Organizations should separately consult the relevant laws and regulations across all jurisdictions in which they operate.

In addition, the process of harmonizing the large set of recommendations into a manageable size inherently results in the loss of information. Although we have taken steps to validate that our framework captures the most salient content and accurately reflects the underlying guidance, our recommendations will lack some of the specificity provided in the original reports. To help address this issue, we provide a crosswalk document to supplement this report. This resource provides a map between the recommendations in our harmonized framework and those of the original guidance documents. This information can be found in the report's supplemental materials. Organizations should use the crosswalk as a reference to locate relevant recommendations from other reports. This can enable practitioners to dive deeper into any given topic or recommendation from our framework.

Harmonization Results

Applying our methodology to the corpus of 52 reports, we distilled the 7,741 recommendations into a representative set of 258 harmonized recommendations. This framework, which we present later in this report, is organized into five high-level categories and covers 34 distinct topics. In this section, we present the results of the harmonization process and provide evidence to support the validation of our methods. First, we illustrate the results of the clustering analysis. Second, we examine the results of the qualitative analysis used to synthesize the harmonized set of recommendations.

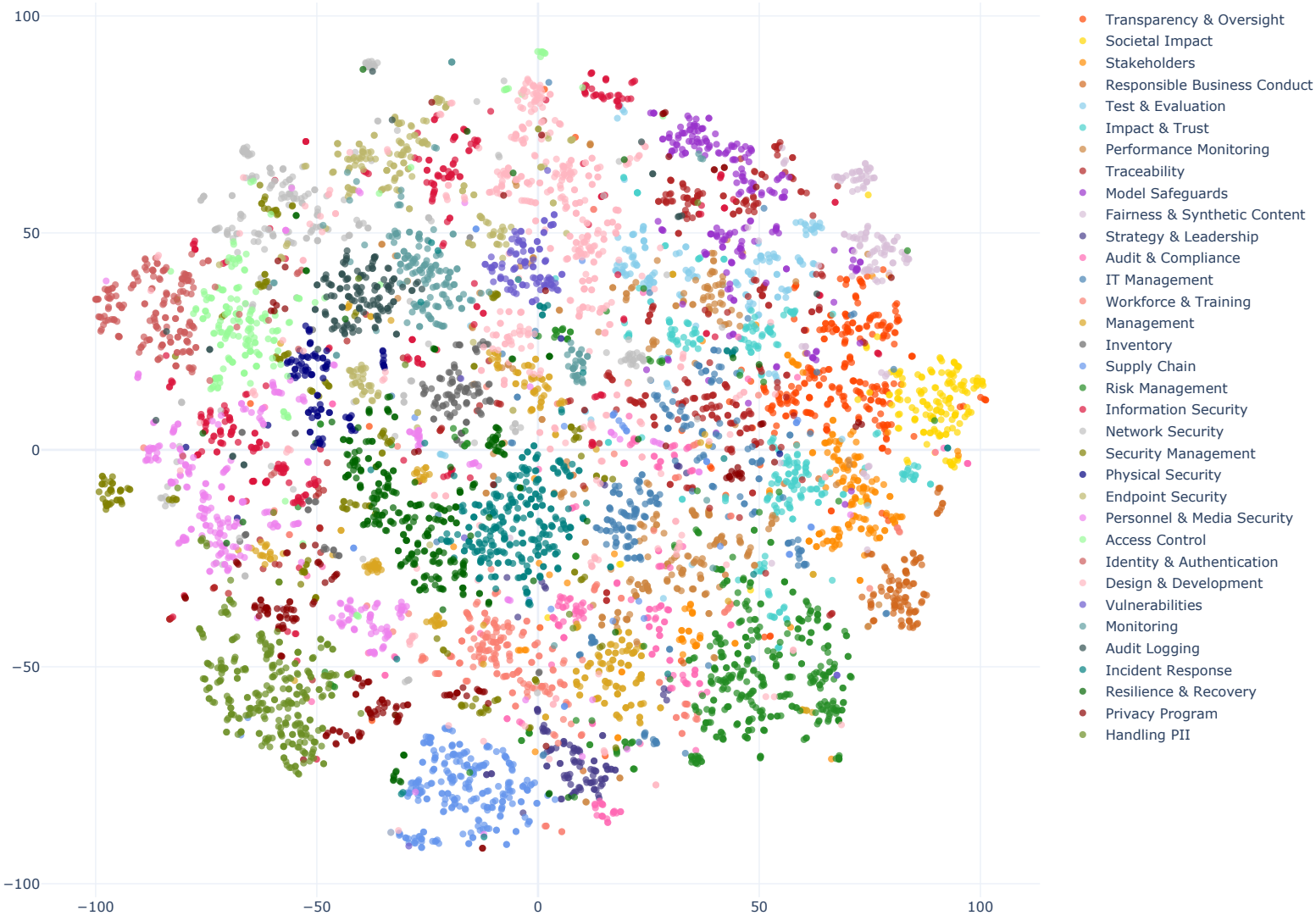
Clustering

We derived the two-tier hierarchical structure of our harmonized framework from the results of our clustering analysis. Using these methods, we found that 34 distinct topics naturally emerged from the dataset. These 34 topic clusters are displayed in Figure 1, which projects the recommendations—each represented by a dot and color-coded by topic—into a two-dimensional representation of the embedded recommendation space. The topics cover a wide range of guidance, spanning from high-level organizational practices, such as risk management or audit and compliance, to low-level technical topics, such as access control or audit logging.

The number of recommendations associated with each topic varies widely (μ=227.1, σ=90.9), which can be interpreted as a combination of topic breadth, relative importance, and the availability or applicability of existing guidance to AI systems. The Design & Development, Risk Management, and Incident Response clusters comprised the most recommendations with 485 (6.3%), 444 (5.7%), and 326 (4.2%), respectively. In contrast, the Physical Security, Vulnerabilities, and Responsible Business Conduct clusters had the fewest recommendations with 97 (1.3%), 108 (1.4%), and 110 (1.4%), respectively. A detailed breakdown of the clusters can be found in the appendix.

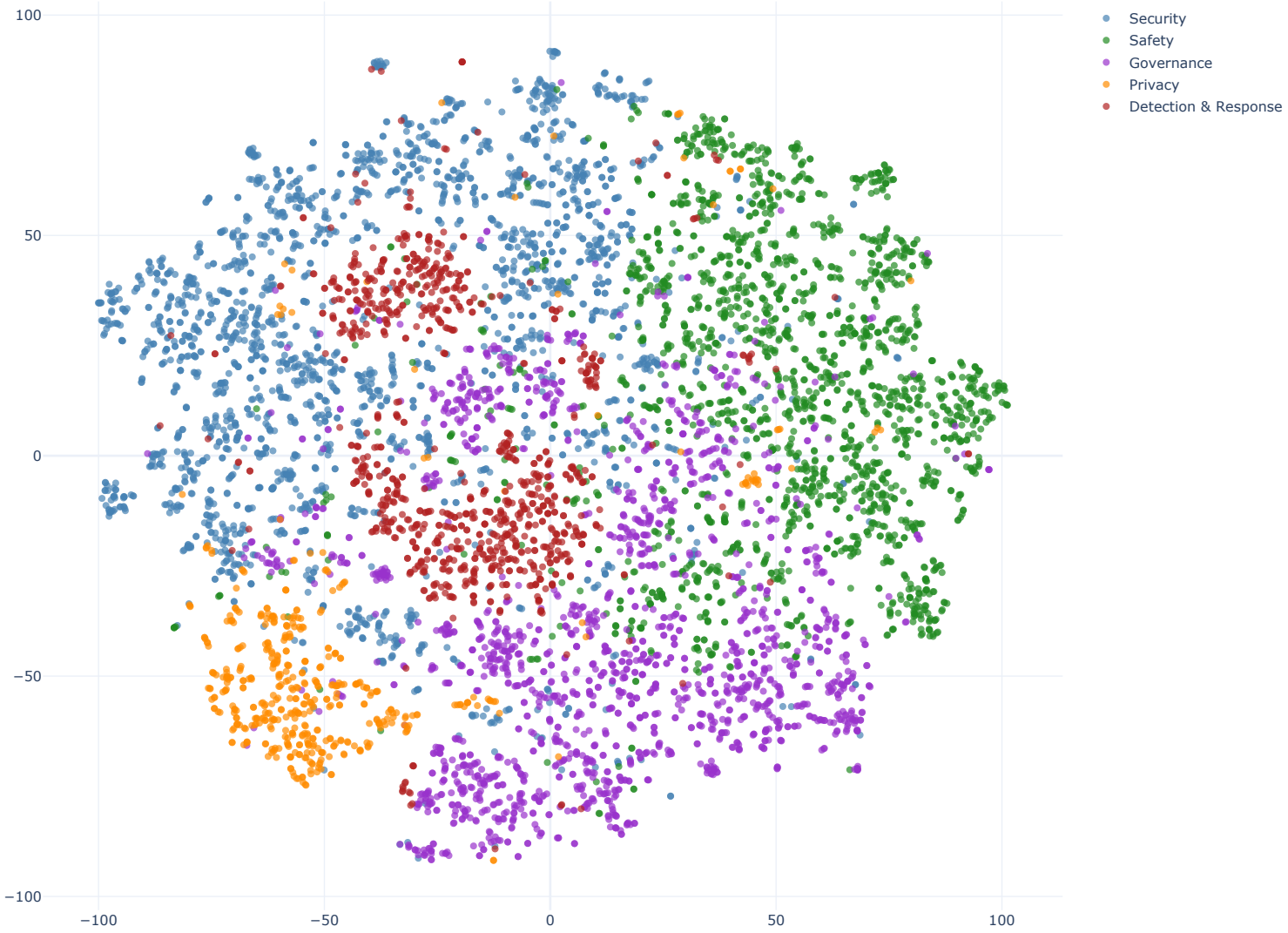
These topics are further grouped into five overarching categories: Governance, Safety, Security, Privacy, and Detection & Response. Figure 2 presents the same visualization of the recommendation space but with each recommendation color-coded by category rather than topic.

Figure 1: Recommendation Grouped and Color-Coded by the 34 Topic Areas



Source: CSET.

Figure 2: Recommendations Grouped and Color-Coded by the Five Categories



Source: CSET

The proximity of the various topic and category clusters in both figures also provides insight into how the different functions within the organization relate to one another. Evident in both figures is the central role that Detection & Response plays in relation to the other categories. This reflects the need for organizations to continually monitor for security incidents, privacy breaches, safety violations, and other sources of risk. In addition, this represents the feedback loop between the lessons learned in response to an incident and the improvement of existing safety, security, privacy, and risk management practices. Each of the topics within the Detection & Response category revolve around the Inventory cluster, indicating the importance of effectively identifying and tracking organizational assets to reliably detect potential incidents involving them.

We can observe other closely related functions along the boundaries of the other categories. There are many commonalities between aspects of Security and Privacy, particularly in maintaining the confidentiality of information. This is reflected in the proximity of the Information Security and Personnel & Media Security topic clusters and those included in the Privacy category. We also observe a substantial overlap between the Safety and Governance categories involving the Stakeholders and Impact & Trust topics on the Safety side and Responsible Business Conduct and Risk Management on the Governance side. We find a similar relationship between Personnel & Media Security within the Security category and Workforce & Training within the Governance category. In this case, both topics focus on different aspects of policies related to an organization's employees. At the intersection of Security and Safety we find the Model Safeguards cluster, an area where we observe many traditional cybersecurity concepts—such as red-teaming—being adapted for AI safety. Finally, between the Privacy, Security and Governance categories we find the Supply Chain cluster. This topic incorporates recommendations related to the data supply chain, often the focus of privacy concerns, and the software and hardware supply chain, which typically falls under the cybersecurity domain.

Framework Validation

We synthesized the harmonized set of recommendations by applying our qualitative coding approach to each of the 34 topic areas, distilling the 7,741 recommendations in the original corpus down to 258 recommended practices. This represents a much more manageable amount of information for organizations to digest, helping to address the information overload problem. Through the process of qualitative coding we identified the most salient recommendations and reduced a substantial amount of redundancy. However, there is also an inherent volume of information lost during the harmonization

process as we compressed the size of the recommendation set. To validate our results, we evaluated the accuracy, representativeness, and completeness of the harmonized recommendations in relation to the original corpus.

To conduct this validation, we generated vector embeddings for each of the harmonized recommendations using the same process described in the Methods section. We then projected the recommendations into the same embedding space as the original corpus. The results are displayed in Figure 3, with the harmonized recommendations highlighted in bold.

Figure 3: Projection of the Harmonized Recommendations (Bolded) into the Same Recommendation Space

Source: CSET.

Accuracy: In this context, accuracy pertains to how similar the harmonized recommendations are to those in the original corpus. As illustrated in Figure 3, the harmonized recommendations neatly align with the original clusters in the embedded recommendation space. This is evidenced by the bolded dots (the harmonized recommendations) being located near lighter dots (the original recommendations) of the same color. This indicates that the harmonized recommendations accurately reflect the content of the original recommendations

Representativeness: For our purposes, representativeness reflects how well the set of harmonized recommendations in each topic cluster represents the broader set of original recommendations. This is evidenced by the harmonized recommendations being well distributed across their corresponding clusters. While this is evident in Figure 3 (above), it is further illustrated in Figure 4 (below) which shows three example clusters—Monitoring, Risk Management, and Information Security—that demonstrate good within-cluster representation.

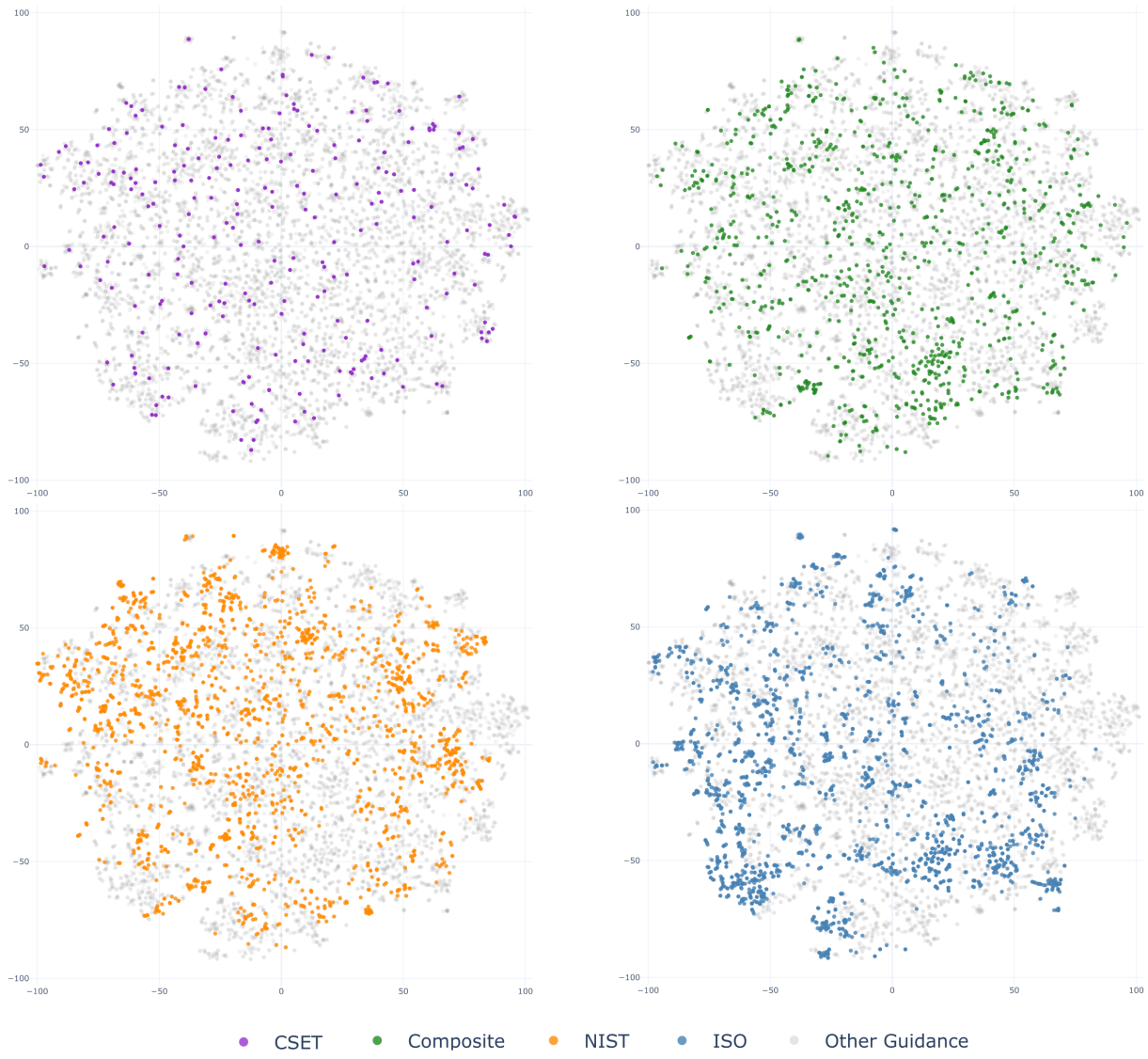
Figure 4: Example of the Harmonized Recommendations (Bolded) Overlaid with That of the Original Corpus

Source: CSET.

Completeness: In this case, completeness refers to how well the harmonized set of recommendations covers the breadth of the recommendation space. With 258 recommendations, our harmonized framework is larger than most of the 52 individual reports (μ=148.9 σ=208.4) in our corpus. That is because the harmonized set covers a much wider breadth of topics than any of the original guidance documents. To get comparable coverage of the recommendation space an organization would need to incorporate guidance from a composite set of seven different reports, comprising a total of 946 recommendations. Figure 5 demonstrates the coverage of the harmonized

set of recommendations—which is evenly distributed across the recommendation space—in comparison to the coverage of the aforementioned composite set of seven reports, the primary NIST frameworks, and the ISO standards. Overall, the harmonized set achieves similar, if not better, coverage with many fewer recommendations.

Figure 5: Coverage of the Harmonized, Composite, NIST, and ISO Recommendations



Source: CSET.

Harmonized Set: 258 recommendations, 1 report

This is the set of recommendations presented in this report. These were added by embedding the text of the harmonized recommendations as described in the Methods section and projecting them into the same recommendation space.

Composite Set: 946 recommendations, 7 reports

This is the minimum set of reports needed to achieve similar coverage to the harmonized set. This set includes recommendations from the NIST AI RMF, CSF, and Privacy Framework; ISO/IEC 27001; CIS Critical Security Controls; the UK NCSC's Principles for the Security of Machine Learning; and UC Berkeley CLTC's Taxonomy of Trustworthiness for Artificial Intelligence.

NIST Reports: 1,920 recommendations, 6 reports

This is the coverage of the recommendation space if only using the primary NIST frameworks and special publications: AI RMF, AI RMF Playbook, CSF, Privacy Framework, SSDF, and SP 800-53.

ISO Standards: 2,106 recommendations, 6 reports

This is the coverage of the recommendation space if only using the following ISO standards: ISO/IEC 23894, 27001, 27002, 27701, 31000, and 42001.

CSET's Harmonized AI Framework

Our harmonized framework consists of 258 recommendations for how an organization should govern, manage, and protect its technology—and how to integrate the management of AI systems into its existing practices. Alongside each recommendation in our framework, we provide an AI Score that indicates what percentage (normalized) of the source recommendations originate from AI-specific reports. Use this score to better understand where new AI guidance overlays with existing organizational practices. A higher score indicates that the recommendation was derived primarily from AI-specific guidance, while a lower score means it comes mostly from the broader cybersecurity, privacy, or risk management literature. We intend this framework to be used as a resource for practitioners and policymakers alike. To this end, we envision our harmonized framework serving the following purposes:

For practitioners, this framework outlines a comprehensive approach for your organization to manage its technological assets in the age of AI. Use this framework to understand how the practices across a wide variety of disciplines fit together, find recommendations that are pertinent to specific topics of interest, and prioritize the most salient best practices. This framework serves as a guide to help organizations plan the implementation of their technology, but should not be treated as a checklist. Adapt these recommendations to the specific needs of your organization. Look to forthcoming CSET reports for further guidance on implementing and tailoring this framework.

For policymakers, the breadth of topics covered in this framework provides insight into how much is already being asked, at least voluntarily, of organizations. Use this framework to better understand the approach that organizations are taking to develop and deploy AI systems, identify practices that are most important to the public interest, and assess how ecosystem-wide infrastructure and policies may connect to and support these efforts. In considering potential regulation or legislation, use this framework to identify and help assess the potential impact on an organization's day-to-day operations.

To facilitate searching for relevant guidance, the framework is organized into 34 topics areas and grouped into five overarching categories as outlined below:

Governance: Defining and implementing the overarching organizational strategy for managing technology and its associated risks	
Topics:	*Strategy & Leadership, Management, Risk Management, IT Management, Supply Chain, Workforce & Training, Inventory, Audit & Compliance*

Safety: Responsibly developing and evaluating the organization's technology, assessing the impact it has on society, and engaging with stakeholders to foster trust	
Topics:	*Responsible Business Conduct, Stakeholders, Societal Impact, Impact & Trust, Fairness & Synthetic Content, Test & Evaluation, Performance Monitoring, Traceability, Transparency & Oversight, Model Safeguards*

Security: Developing and deploying secure systems, managing access to facilities and assets, and implementing security controls	
Topics:	*Security Management, Design & Development, Vulnerabilities, Identity & Authentication, Access Control, Network Security, Information Security, Endpoint Security, Personnel & Media Security, Physical Security*

Privacy: Managing data, particularly personally identifiable information (PII), and protecting the privacy and confidentiality of data throughout its life cycle	
Topics:	*Privacy Program, Handling PII*

Detection & Response: Identifying threats and incidents, responding when these events occur, and building greater operational continuity	
Topics:	*Audit Logging, Monitoring, Incident Response, Resilience & Recovery*

Governance

The governance section focuses on defining and implementing the overarching organizational strategy for managing technology and its associated risks.

Strategy & Leadership (SL)			AI Score
1	Organizational Strategy	Define an organization-wide strategy for managing technology. Develop policies that address cybersecurity, safety, privacy, and risk management. Communicate and enforce policies across the organization.	
2	Leadership	Designate executive roles (e.g., CIO, CTO, CISO) to oversee the organization's technology. Ensure that the direction from leadership at the board level is translated into effective organizational practices. Hold leadership accountable for performance.	
3	Oversight	Conduct regular management reviews of each business unit's governance and risk management activities. Revise these policies as the organization's technology and cybersecurity risk, and risk tolerance, change.	
4	Integrated Risk Management	Integrate the range of risk management activities (AI, cybersecurity, privacy, supply chain, etc.) into the enterprise risk management program. Coordinate and align these activities with each other and with those of external partners.	
5	Culture	Ensure that leadership demonstrates a commitment to safety, security, privacy, and accountability. Communicate openly about risks and empower individuals at all levels to report issues and concerns.	
Management (MG)			**AI Score**
1	Context and Requirements	Identify the safety, security and privacy obligations of the organization to its customers and stakeholders, including all legal, statutory, regulatory, or contractual requirements. Use these requirements to define objectives and scope management activities.	
2	Strategy and Objectives	Develop a strategy and objectives to meet the requirements of the organization. Ensure management is committed to these objectives and communicates priorities to personnel.	
3	Management System	Establish a management system that implements mechanisms and activities to achieve the organization's objectives. Provide adequate resources to enable effective management.	

4	Policies and Procedures	Establish, document, approve, communicate, apply, evaluate, and maintain policies and procedures that cover the range of safety, security, privacy, and risk management activities.	7
5	Documentation	Document policies and procedures. Keep documentation of activities, decisions, and outcomes related to policies. Make documented information available for auditing.	56
6	Data Management and Retention	Manage and protect documented information and audit data. Establish a data retention policy and destroy data in accordance with the retention period.	34
7	Management Practices	Establish change management, configuration management, and exception management practices.	22
8	Review and Improve	Regularly evaluate the effectiveness of policies, procedures, and the management system. Conduct regular management reviews. Continuously improve and update the management system.	64
Risk Management (RM)			**AI Score**
1	Establishing Context	Establish the context of the risk management process based on the internal and external environment in which the organization operates. Collect input from stakeholders and experts to support this process.	53
2	Strategy and Tolerance	Establish a strategy for risk management. Define organizational objectives and risk tolerance.	56
3	Process, Roles, and Culture	Integrate risk management into organizational decision-making and culture. Communicate roles and responsibilities for risk management to personnel.	54
4	Risk Identification	Identify threats, vulnerabilities, and risks to the organization.	40
5	Risk Assessment and Analysis	Assess the likelihood and magnitude of risks, quantitatively or qualitatively, accounting for uncertainty.	45
6	Risk Prioritization	Prioritize risks based on the criticality of assets affected and impact on organizational mission.	40
7	Risk Response	Develop and implement a risk treatment plan, applying appropriate controls to mitigate risk. Respond to unanticipated risks when they arise.	51
8	Risk Tracking	Track mitigated and unmitigated (residual) risk. Monitor the effectiveness of risk treatments.	64

9	Continual Improvement	Regularly test, review, and improve the risk management process. Incorporate lessons learned from unanticipated risks when they arise.	
10	Transparency and Communication	Communicate information about risk, the organization's actions to mitigate them, and the effectiveness of those actions to internal and external stakeholders.	
IT Management (IT)			**AI Score**
1	Mission, Strategy, and Alignment	Articulate the organization's mission and develop a strategy to achieve its objectives. Align IT initiatives and systems with organizational goals.	
2	Context and Dependencies	Identify the internal and external context of the organization, paying particular attention to advances in technology. Identify the organization's dependencies and stakeholders' dependencies on the organization's services.	
3	IT and Risk Management	Establish ownership over IT infrastructure within the organization. Assign responsibility for implementing the IT strategy, managing the IT portfolio, mitigating IT-related risks, and overseeing the life cycle of IT systems.	
4	Architecture	Establish a common IT architecture, manage enterprise architecture services, and determine how new systems will be deployed and integrated into the existing architecture.	
5	Resource Allocation	Budget and allocate resources (e.g., money, people, technology) across IT projects and programs. Prioritize investments that support the organization's mission and optimize expected net benefits. Account for resource constraints.	
6	Project and Program Management	Maintain a standard process for managing IT projects and programs. Define life cycle stages and ensure proper review and approval of work at each stage.	
7	Design and Development	Design and develop systems responsibly and in alignment with organizational values. Clearly define the system's purpose and requirements. Develop the system to meet those requirements.	
8	Assessment and QA	Evaluate developed and acquired systems against prespecified release criteria prior to deployment. Ensure the system passes the quality assurance process. Use the results to make a go/no-go decision and obtain authorization for deployment.	

9	Deployment and Migration	Establish a deployment plan and adapt based on pre-deployment evaluations. Follow organization-approved processes for deployment. Use pilots and staggered releases to limit risk.	86
10	Operation and Decommissioning	Manage the system post-deployment in accordance with service-level agreements. Establish a measurable baseline and monitor performance for deviations. Assess the impact of the system's operation and eventual decommissioning. Report results to stakeholders.	62
Supply Chain (SC)			**AI Score**
1	Managing and Coordinating	Manage relationships with suppliers. Establish procedures to acquire, use, manage, and exit third-party services. Coordinate roles and responsibilities between the organization and suppliers.	26
2	Mapping the Supply Chain	Map the third-party products, components, and services that the organization depends on throughout the supply chain. Identify suppliers and alternative suppliers, prioritizing by criticality.	47
3	Risks and Risk Management	Incorporate supply chain risks into enterprise risks management. Ensure third-party risks are included as a part of risk identification, assessment, and treatment activities.	39
4	Supply Chain Security	Raise awareness and improve the resilience of supply chain security. Ensure the organization meets its own security responsibilities as a supplier and consumer. Require the same of its suppliers.	58
5	Procurement Plans and Due Diligence	Establish procurement plans to evaluate and select services from a range of options. Vet potential suppliers and conduct due diligence prior to acquisition. Collect evidence that suppliers and services meet the organization's standards.	39
6	Contracts and Requirements	Establish contractual obligations that require third-parties to implement specified security, privacy, risk management, audit, and compliance practices. Include provisions to verify those obligations are met.	20
7	Monitoring and Assessment	Assess the practices of third-party organizations and their continued ability to comply with the terms of their contract. Continuously monitor third-party services and products for changes, deviations, or failures in meeting obligations.	17

Workforce & Training (WF)			AI Score
1	Training and Awareness	Work with personnel to develop their risk management, privacy, and cybersecurity skills. Provide general awareness and role-specific training. Conduct regular competency and knowledge checks to ensure ongoing compliance.	30
2	Roles and Responsibilities	Establish clear internal governance structures with delineated roles and responsibilities for risk management, monitoring and response, safety, security, and privacy functions.	51
3	Resource Allocation	Provide personnel and human resources with adequate resources to implement and carry out organizational processes, policies, and controls.	19
4	Policy and Enforcement	Establish, communicate, and enforce organizational policies. Ensure that personnel understand their responsibilities in upholding these policies and hold individuals accountable for doing so.	35
5	Culture	Foster a positive safety, cybersecurity, privacy, and risk-aware culture. Demonstrate a commitment to inclusivity, collaboration, and ethical values. Hire and retain competent personnel who will help further these goals.	43
Inventory (IV)			**AI Score**
1	Inventory	Maintain an up-to-date inventory of all organizational assets including systems, components, hardware, software, data, devices, users, and accounts. Include third-party assets.	25
2	Discovery and Tracking	Use tools for the automated discovery and tracking of organizational assets. Use these tools to automatically update inventory information.	0
3	Mapping	Maintain an up-to-date mapping of the organization's networks and data flows.	0
4	Asset Management	Actively manage all assets and their configuration, throughout their life cycle. Classify assets according to criticality and sensitivity. Protect assets according to their classification.	32
5	Ownership and Responsibility	Assign and document ownership of assets within the organization. Designate responsibility for managing assets and maintaining inventories.	34

Audit & Compliance (AU)			AI Score
1	Audit Oversight	Create governance structures within the organization that includes establishing independent risk management and oversight functions.	
2	Compliance	Ensure compliance with all relevant legal, regulatory, and contractual requirements. Monitor changes to the legal and regulatory landscape. Adhere to industry standards and best practices. Regularly review to ensure continued compliance.	
3	Audit	Establish an independent audit and assurance function within the organization. Conduct audits regularly. Define objectives, criteria, and scope for each audit.	
4	Adherence to Policies	Ensure that personnel within the organization adhere to established policies and that policies are implemented as intended. Design policies to be practical and usable.	
5	Third-Party Systems and Use	Establish policies that address the procurement, testing, and use of third-party systems and models. Oversee third-party use of organizational systems to prevent misuse.	
6	Culture and Collaboration	Promote an organizational culture that prioritizes safety and risk mitigation. Encourage collaboration and communication between teams, both internally and externally.	
7	Communication and Disclosure	Communicate the organization's commitment to governance, risk management, and compliance to stakeholders. Share policies, practices, and outcomes related to these initiatives publicly.	

Safety

The safety category covers how organizations should responsibly develop and evaluate technology, assess the impact on society, and engage with stakeholders.

Responsible Business Conduct (RC)			AI Score
1	Responsible Business Conduct	Develop responsible business conduct (RBC) policies. Publicly publish RBC policies and activities. Assign oversight of RBC issues to senior management.	
2	Due Diligence	Conduct due diligence prior to establishing business or contractual relationships. Regularly assess RBC risks and human rights impacts arising from business relationships, incorporating feedback from stakeholders.	
3	Relationship Management	Build relationships with business partners to promote the adoption of RBC practices, specify exceptions contractually when feasible. Work with partners to plan for and mitigate adverse events when they arise.	
4	Accountability	Hold the organization, vendors, and suppliers accountable for maintaining RBC practices. Establish complaint procedures for workers. Identify RBC violations and their causes. Apply appropriate remediation.	
5	Public Reporting	Communicate RBC and due diligence practices to the public. Report the results of due diligence assessments, including identified RBC risks and violations.	
6	Remediation	Stop activities causing or contributing to adverse RBC events. Identify and engage with impacted individuals or communities. Cooperate in good faith with legal, judicial, or other remediation mechanisms to provide appropriate compensation to affected parties.	
Stakeholders (ST)			**AI Score**
1	Stakeholder Engagement	Identify internal and external stakeholders that may be impacted by the organization's products or services, directly or indirectly. Involve stakeholders and their input in all stages of a system's life cycle.	
2	Communication	Establish clear communication channels with stakeholders. Communicate information about risks and mitigation efforts to build trust. Provide mechanisms to enable regular communication and foster dialogue with stakeholders.	

3	Feedback Mechanisms	Provide feedback mechanisms to identify stakeholders' priorities and concerns, incorporate input into internal decision-making, identify negative impacts, and evaluate the effectiveness of mitigations.	
4	Inclusive Development	Establish diverse and interdisciplinary development teams. Supplement organizational diversity by seeking the input and knowledge of a diverse set of stakeholders and experts in the development process.	
5	Competence and Expertise	Ensure that team members have the appropriate knowledge and competence required for safety, security, and risk management activities. Seek out external domain expertise as needed.	
6	Socio-technical Evaluation	Employ human-centered design principles in developing systems. Test systems in collaboration with socio-technical, human factors, user interaction/user experience (UI/UX), and human-computer interaction (HCI) experts.	
Societal Impact (SI)			**AI Score**
1	Human-Centric Development	Promote ethical and human-centric development of technology that benefits society. Collaborate with different industries, civil society, and academia to foster ethical research and develop shared best practices.	
2	Equity, Inclusion, and Access	Develop and deploy technology that promotes fairness. Ensure that systems combat stereotyping and discrimination. Promote widespread and equitable access to the organization's tools and services.	
3	Workforce and Economy	Contribute to innovation that benefits the whole of society. Promote a fair and competitive business environment. Account for the impact that transformative technology, such as AI, can have on the workforce. Support education, training, and re-skilling efforts.	
4	Collaborative Governance	Involve external stakeholders in internal governance efforts. Participate in collaborative initiatives to develop norms, share knowledge, and advance safety and security across the ecosystem.	
5	Societal and Global Stability	Provide tools and services that improve, rather than subvert, social and civic processes. Ensure responsible and controlled use in military domains. Promote research on health, mental health, and safety impacts of technology.	

6	Sustainability and Environment	Develop sustainable and environmentally friendly technology. Promote sustainable business practices. Responsibly manage the organization's use of natural resources, energy, and production of pollutants.	100
7	Global Governance	Promote international cooperation and collaboration. Support international governance initiatives, standards development, and research. Respect international law. Contribute to efforts in tackling global challenges.	100
Impact & Trust (IM)			**AI Score**
1	Impact Assessments	Conduct regular impact assessments to identify and measure the impact of potential failures, disruptions, or harmful output of a system. Account for the nature of the system, its operating environment, and involved stakeholders.	99
2	Documentation and Collaboration	Document the impact assessment process and the risks identified. Document the intended purpose of a system and its potential benefits. Collaborate with third parties to establish context-specific auditing mechanisms to evaluate real-world impacts.	100
3	Trustworthy System Design	Promote the development of trustworthy systems, particularly those that are AI- or ML-based. Obtain the requisite talent to build trustworthy systems and establish robust testing, evaluation, verification, and validation (TEVV) practices.	98
4	Robustness	Develop systems that are robust against failures, misuse, and malicious attacks. Implement measures to reduce safety and security risks. Evaluate these capabilities under normal and adverse conditions.	100
5	Ethical and Societal Implications	Assess and document the potential societal impacts of the organization's systems on human rights, physical and mental health, privacy, democratic values, and societal well-being—particularly when AI is involved.	100
6	Environment and Sustainability	Assess the organization's environmental and ecological footprint, including the energy and water consumption and carbon emissions related to its use of technology. These can be particularly acute for AI systems.	100
7	Performance Trade-offs	Assess, quantitatively and qualitatively, the potential benefits and costs of risks and impacts. Articulate and analyze the trade-offs between trustworthy characteristics and performance.	100

8	Ensuring Long-term Trustworthiness	Continue to evaluate the trustworthy characteristics and measure downstream impacts for deployed systems. Maintain the system and regularly reassess its impacts over time.	100
Fairness & Synthetic Content (FS)			**AI Score**
1	Bias in Datasets	Scrutinize datasets for bias, including distributional differences across subgroups; lack of completeness, representativeness, or balance in the data; features or proxies that convey sensitive or demographic information; and embedded historical, systemic, or human-cognitive bias.	100
2	Detecting Bias	Identify fairness metrics and benchmarks to monitor bias in model performance. Conduct fairness assessments and disaggregated or subgroup analysis to identify within-group and intersectional disparities.	100
3	Mitigating Bias	Incorporate activities to mitigate bias into the organization's development, deployment, and operation of its systems and models. Be transparent to stakeholders about the sources of training data, potential bias, and related ethical considerations.	100
4	Post-deployment Monitoring	Monitor and prevent or mitigate bias, skewed responses, and the generation of harmful or manipulative output from deployed systems. Use structured feedback mechanisms to help identify these issues.	100
5	Synthetic Content	Disclose the use and distribution of synthetic content and media. Employ provenance methods such as watermarking, cryptography, and steganography. Obtain informed consent from and maintain attribution of the creators, subjects, and content sources of synthetic media.	100
Test & Evaluation (TE)			**AI Score**
1	Policy and Planning	Establish a testing strategy that includes acceptance criteria for new systems and models. Develop a plan for TEVV activities.	80
2	Life cycle Cadence	Define the frequency and specific life cycle stages at which TEVV activities occur. Conduct regular testing both before and after deployment, including when any changes are implemented.	59
3	Testing and Replication	Ensure the reproducibility of system outputs, model training, and testing results. Be able to replicate the results of third-party testing. Record testing results and make them available using replication files.	95

4	Types of Testing	Evaluate system's validity, robustness, repeatability, and domain fit. Verify system changes. Assess third-party claims. Conduct regular red-teaming, penetration, and security testing.	
5	Review Results	Review the results of the TEVV process and resolve issues within a defined time period. Regularly reevaluate the effectiveness of TEVV metrics and processes.	
6	Documentation	Be transparent about the TEVV process and its results. Document TEVV process, including test sets, assumptions, metrics, testing procedures, techniques, and results.	
7	Expert Involvement	Engage subject matter and domain experts in the TEVV process. Have external experts conduct testing and validate results of internal testing. Undertake independent acceptance testing.	
8	Automating Testing	Implement automated testing and validation mechanisms that can verify against known facts or data. Use automated methods or generate synthetic data to expand the comprehensiveness of testing.	
Performance Monitoring (PM)			**AI Score**
1	Monitoring Performance	Determine what technical and business metrics should be measured and monitored over the course of the system's life cycle. Conduct continuous monitoring and regular validation of system performance.	
2	Performance Drift and Misalignment	Measure the model's performance for drift, misalignment, and behavior change. Conduct regular health checks to ensure the model continues to align with the organizations' values and risk tolerance. Review when new versions are deployed.	
3	Continuous Reassessment	Conduct regular risk and impact assessments, evaluations, red-teaming exercises, and penetration testing. Continuously reassess the effectiveness of the tests and metrics being used.	
4	Corrective Action	Take corrective action when an issue, noncompliance, or nonconformity is identified. Assess the effectiveness of the corrective action taken. Update response and recovery strategies as necessary.	
5	Oversight	Continually identify improvements from evaluation and monitoring activities. Validate that these activities provide sufficient information for audit, compliance, and oversight.	

Traceability (TR)			AI Score
1	Data Provenance and Lineage	Put in place methods to track the data lineage and provenance of a system. Maintain metadata records of the data's origin, associated labels or categories, processing, change history, use limitations, and retention policy.	
2	Version Control and Change Tracking	Implement a version control system to manage changes made to data, datasets, source code, system artifacts, and model weights. Store metadata about each change, the reasons for it, and how it was implemented, tested, and deployed.	
3	System and Data Documentation	Document a system's intended use, risks, capabilities, and limitations. Record system design, development, testing, and deployment details. Document data, data elements, and processing.	
4	Auditability	Create an audit trail that can trace the system's outputs back to the rules, algorithms, and data that was used to generate it. Be able to trace data within the organization from collection to disposal.	
5	Data Quality and Validation	Implement a systematic approach to data quality. Improve the quality, completeness, suitability, and representativeness of data used to train models. Validate and monitor the quality of data over time.	
6	Evaluation Documentation	Maintain a systematic record of measurement and evaluation results. This includes the output of tests and materials to reproduce them, performance metrics, and resource utilization.	
Transparency & Oversight (TO)			**AI Score**
1	User Awareness and Communication	Provide users of a system with documented instructions, guidance, and training on its proper use. Convey information on the system's risks and limitations. Build informative alerts and notifications into the operation of the system.	
2	Explainability and Interpretability	Establish transparency, explainability, and contestability (TEC) requirements for system development and use. Where possible, provide explanations to users on how decisions or outputs were reached.	
3	Human Oversight and Accountability	Meaningfully incorporate human oversight and agency into the design of models and systems. Ensure that a human-in-the-loop remains accountable for system output and in control of its operation.	

4	Data and Safeguard Transparency	Put in place mechanisms to flag issues of bias, harmful output, poor performance, and misuse. Provide transparency around the implementation of these safeguards without violating their integrity.	98
5	Alignment with Human Values	Carry out transparent self-assessments of how organizational policies and technologies align with human values, standards, regulatory frameworks, and the rule of law.	100
6	Mechanisms and Documentation	Establish formal mechanisms to build transparent practices into the organization's development and use of technology. Produce transparency reports and model cards to disclose details about the development or use of AI models.	100
7	Public Accountability	Hold the organization accountable to the public, providing transparency and protecting consumer rights. Engage with stakeholders to ensure that harms caused by the organization or its systems are adequately redressed.	100
Model Safeguards (SG)			**AI Score**
1	Fail-safes	Develop systems to identify and handle out-of-distribution input, low-confidence predictions, and high uncertainty situations in which failures are likely to occur. Employ fail-safes such as deferring to a human-in-the-loop.	100
2	Mitigating Data Risks	Use trusted data labeling and data sources for model training. Assess datasets for potential bias, data quality issues, and signs of poisoning or tampering. Employ training techniques, such as using adversarial examples, to improve model robustness.	100
3	Model Security	Protect models against security threats including adversarial, poisoning, out-of-distribution, model inversion, membership inference, and model extraction attacks. Harden access points, such as application programming interfaces (APIs), and scrutinize inputs and outputs for anomalies.	100
4	Evaluating Performance	Analyze system performance for model degradation, data drift, anomalous behavior, and emergent capabilities. Track key metrics and establish regular benchmarking. Systematically review and report results.	100
5	Model Supply Chain	Source assets that are used in the development of AI systems (e.g., data, libraries, software, hardware, pretrained models) from verified and trustworthy sources. Document sources using bill of materials (BOM) or model cards.	100

6	Continual Learning	Deploy safeguards to sanitize new data used by AI systems for continual learning. Scrutinize changes in model behavior as these systems can be more susceptible to poisoning and adversarial attacks.	

Security

The security section provides guidance on how the organization should develop and deploy secure assets, manage access to facilities and systems, and implement security controls.

Security Management (SM)			AI Score
1	Security and Privacy Program	Establish a program to manage security and privacy across the organization. Communicate the strategy, related policies, and responsibilities to personnel. Enforce and regularly evaluate the program's policies.	0
2	Implementing Security	Design enterprise architecture to be secure. Implement processes and controls to protect organizational assets (e.g., hardware, software, data, systems, and networks).	0
3	Essential Function	Identify and prioritize protecting the data, networks, and information systems supporting the essential function of the organization.	0
4	Encryption and Key Management	Employ accepted methods of encryption to secure assets. Manage and protect the creation, distribution, use, storage, and destruction of cryptographic keys.	0
5	Security Boundaries	Establish physical and logical boundaries at the organization's perimeter and between segregated security domains within the organization. Implement protections at those boundaries.	35
6	Physical and Logical Access	Manage and monitor physical and logical access across boundaries and to assets within those boundaries.	0
7	Information Flows and Transfers	Map and control information flow across security domains in accordance with applicable laws and regulations. Closely control the transfer of personal, sensitive, or classified information.	0
8	Data Management	Identify, catalog, and track the organization's data and where it is stored. Manage data with respect to security, privacy, and applicable laws throughout the data life cycle. Retain data backups.	0
Design & Development (DD)			**AI Score**
1	Policies and Procedures	Establish organization-wide processes to promote secure design and development. Require developers to use secure coding practices. Ensure the organization maintains the capability to develop and support selected technologies.	36

2	Threat Modeling	Conduct threat modeling and attack surface mapping as a part of system design. Ensure the development team is aware of the organization's threat landscape.	
3	Secure Design	Adopt a secure-by-design approach. Implement security design principles. Assess how the system will interact with other IT infrastructure. Ensure that the design specification is consistent with the organization's security and privacy architecture.	
4	Secure Development	Employ secure software development (SSD) practices across the system development life cycle (SDLC). Maintain separate and secure development, test, and deployment environments.	
5	Security Controls	Implement information and network security controls at all stages of the SDLC. Enforce access control and usage restrictions. Employ change control and validation processes to prevent unauthorized changes to system components.	
6	Testing	Require developers, internal or external, to conduct static and dynamic application security testing (SAST and DAST). Commission independent assessments to validate testing.	
7	Reviewing	Establish review processes for both manual and automated review of system design, code, and security processes.	
8	Secure Deployment	Prioritize secure deployment practices. Use staged release and blue-green deployment strategies. Automate deployment mechanisms, incorporating tracking and approval workflows.	
9	Baseline Configurations	Create and maintain common secure baseline configurations and templates. Ensure the configurations incorporate security principles.	
10	Documentation	Identify, document, and publish organization-wide common controls and configurations for system development. Document all security requirements and require developers to demonstrate that system implementation meets top-level specifications.	
Vulnerabilities (VN)			**AI Score**
1	Reporting Processes	Create mechanisms and incentives for internal and external parties to report the existence of bugs and vulnerabilities. Report relevant information about vulnerabilities and patches to stakeholders.	
2	Secure Development	Reduce vulnerabilities by following secure software development practices and conducting vulnerability detection. Obtain security assurances from third-party providers. Only use trusted libraries and components.	

3	Prioritization	Triage reported vulnerabilities to determine their validity, assess the scope of affected systems, categorize the severity of impacts, identify affected stakeholders, and analyze response options. Establish a process to prioritize vulnerabilities.	21
4	Detection	Detect vulnerabilities by monitoring CVEs, analyzing software, and conducting vulnerability scans. Identify unauthorized or out-of-date components. Correlate results from multiple sources and scans.	39
5	Patching	Patch or otherwise mitigate known vulnerabilities in a timely manner. Proactively fix similar vulnerabilities in other software or systems.	32
6	Processes	Establish a technical vulnerability management process to plan and implement risk responses to vulnerabilities. Include a process to manage the risk of vulnerabilities that cannot be patched.	21
7	Testing and Evaluation	Test the effectiveness of a remediation or patch before deployment and verify there are no unintended side effects. If a patch or update is provided by an external partner, verify its authenticity before applying it.	0
Identity & Authentication (IA)			**AI Score**
1	Centralized Identity Management	Establish a centralized system to issue, manage, verify, revoke, and audit identities and credentials.	0
2	Proof and Bind	Proof and verify identities. Bind verified identities to authentication credentials. Avoid shared accounts and credentials.	0
3	Protecting Credentials	Store and transmit credentials securely using approved cryptographic techniques.	0
4	Identification and Authentication	Require identification and authentication before allowing physical or logical access across security boundaries. Reauthenticate access when taking sensitive or privileged actions.	0
5	Security Mechanisms	Employ secure authentication mechanisms that are protected against replay, spoofing, and brute-force attacks. Establish an isolated, trusted communication path for authentication.	0
6	Login	Monitor successful and unsuccessful log ins. Do not provide feedback during log in that may be helpful to an attacker. Notify the user of log in attempts.	0

7	Log off, Lock, and Disconnect	Enable automatic log off, device lock, and session disconnect after a set amount of time or period of inactivity. Invalidate session identifiers and provide notification upon log out.	0
8	Passwords	Train users on and enforce secure password practices. Prohibit weak, commonly used, and reused passwords. Eliminate default passwords. Generate and manage passwords using passwords managers.	9
9	Stronger Authentication	Employ stronger authentication methods, particularly in security-sensitive cases, including multifactor and biometric authentication, single sign-on, authenticators, and public key infrastructure.	11
Access Control (AC)			**AI Score**
1	Access Policy	Establish a policy that defines rules for access control and a process for administering access uniformly across the organization.	11
2	Security Principles	Adhere to the principles of least privilege, least functionality, separation of duties, and zero trust in the design and implementation of the access control policy.	31
3	Types of Access Control	Apply attribute-based access controls (ABAC) if feasible, otherwise apply role-based access controls (RBAC). Consider using dynamic access management in conjunction with either approach.	17
4	Account and Access Management	Manage user and system accounts. Implement procedures to provision, review, modify, and revoke accounts and associated privileges.	9
5	Modifying Access	Modify access rights as conditions and needs change. Obtain authorization when granting new or additional access privileges. Revoke access when no longer required.	5
6	Remote Access	Manage remote access. Implement additional access restrictions, device or configuration requirements, and security measures (e.g., encryption, enhanced authentication) for remote access.	0
7	Privileged Access	Strictly limit and segregate the use of privileged access. Grant privileged access on a temporary basis and only after a more stringent authorization is obtained, such as dual or joint authorization.	0
8	Enforcing Access	Enforce the access control policy and prevent unauthorized access. Override access control mechanisms only in defined circumstances by authorized personnel.	25

9	Monitor and Review Access	Regularly review account and access activity to identify atypical usage and revalidate rights and privileges. Modify and remove access, as necessary, at regular intervals.	
Network Security (NS)			**AI Score**
1	Managing Networks	Manage the organization's network. Protect the integrity and security of the network by controlling access, obfuscating the network from attackers, and employing defense-in-depth techniques.	
2	Segmentation and Separation	Logically or physically segment the network into different security domains. Dynamically isolate systems and areas of the network in response to attacks. Separate security and non-security functions, privileged and non-privileged activity, and conflicting duties.	
3	Data Flows and Controls	Employ firewalls and policy-based content filters at the boundaries between security domains to control connections, access, and information flows.	
4	Connections and Managed Interfaces	Route communication to and from the organization through managed interfaces (proxies, VPNs, etc.). Ensure that devices connecting to the network remotely are trusted and maintain the capability to remotely wipe and track those devices.	
5	Wireless Security	Use cryptographic mechanisms and secure protocols to protect wireless networks. Protect wireless networks from signal-based attacks. Segment wireless networks and consider additional restrictions on their access and use.	
6	Availability	Maintain the availability of networked resources by rate-limiting the number of connections and requests. Optimize systems and load-balance allocated resources. Detect and prevent denial-of-service (DOS) attacks.	
7	Time Synchronization	Synchronize clocks across networked systems and devices using two reliable sources of time. Ensure synchronized time across logging and auditing capabilities.	
8	Minimizing Attack Surface	Develop and enforce a process for decommissioning systems and removing unused components (software, hardware, data, functionality, etc.) to reduce the organization's attack surface and free up resources.	

Information Security (IS)			AI Score
1	Classifying and Categorizing	Implement an organization-wide classification scheme to categorize data and assets by sensitivity or criticality. Define security requirements and processing procedures based on the classification scheme.	
2	Encrypting Data and Communications	Encrypt communications channels. Protect system artifacts (code, model weights) and data (at rest, in transit, in use) using encryption commensurate with its classification.	
3	Integrity and Verification	Use integrity checking mechanisms to verify the authenticity of and prevent tampering with hardware, software, firmware, code, and data. Use cryptographic methods to verify the identities of trusted parties.	
4	Data Quality and Sanitization	Monitor, filter, and sanitize system inputs to prevent incoming attacks and outputs to flag harmful, false, privacy-sensitive, or illegal content. Employ measures to ensure the quality of datasets.	
5	Preventing Data Leaks	Prevent the leakage, exfiltration, and theft of the organization's information and assets. Monitor channels where data leakage can occur. Scan open-source information to identify unauthorized disclosures.	
6	Improving Security Measures	Routinely test security controls and protection mechanisms. Share information about their effectiveness. Automatically update and continuously improve protection technologies.	
7	Auditability	Ensure the auditability of systems. Maintain audit trails and chain of custody to ensure the provenance of data and decisions. When incidents occur, collect and preserve forensic evidence.	
8	Plugins and APIs	Ensure plugins and APIs are implemented securely, following the principle of least functionality. Ensure that only trusted plugins and APIs are used.	
9	Public Release	Designate staff to control the public release of information, materials, and products. Assess the risks of disclosing information and making code or models open source.	
Endpoint Security (ES)			**AI Score**
1	Managing and Tracking Assets	Manage endpoint devices connected to the organization's network. Use bill of materials (BOM) to track hardware (HBOM) and software (SBOM) components.	

2	Unauthorized Components	Prevent the installation and use of unauthorized software (e.g., applications, libraries, code, binary) and hardware components. Detect and remove unauthorized components.	0
3	Unauthorized Changes	Prevent unauthorized changes to source code and the configuration of devices or systems. Prevent privilege escalation and the use of utility programs that can enable unauthorized changes.	11
4	Integrity and Verification	Verify the integrity and security of software and hardware components, particularly those from third-parties. Assess how third-party components will be supported and maintained.	42
5	Malware	Deploy anti-malware protections on devices. Keep repositories of known malware signatures updated. Provide malware training to personnel. Plan for and respond to malware compromises.	0
6	Safeguards	Employ internet and email safeguards including firewalls, spam filtering, blocklists of malicious websites, and secure protocols.	8
7	Maintenance	Regularly maintain hardware and software components, using preventive or predictive maintenance where applicable. Ensure maintenance is done by authorized parties. Log and monitor maintenance activities.	0
8	Updates and Patches	Keep software and hardware up-to-date with patches, updates, and security fixes. Ensure updates are authorized and tested prior to applying them.	40
Personnel & Media Security (MS)			**AI Score**
1	Managing Data and Media	Manage media, and data stored on it, throughout its life cycle. Ensure the secure disposal of media and destruction of data.	22
2	Data and Media Transfer	Control the transfer of media and data, whether physical or digital, across security boundaries. Prevent unauthorized transfers. Protect data in transit.	4
3	Preventing Leakage and Compromise	Employ scanning and sanitization methods to prevent removable media from introducing malware. Prevent information leakage via removable media, electromagnetic signals, eavesdropping, and side channels.	0
4	Background Checks and Suitability	Carry out background checks and screening commensurate with the position being hired for. Complete screening before providing access to systems or data.	0

5	Contractual Agreements	Ensure personnel comply with the obligations stipulated in their employee contract, including access and nondisclosure agreements. Establish acceptable use policies for end users.	
6	Personnel Security	Restrict employee access to certain software, services, websites, and secure areas as required. Enforce secure office practices such as lockable storage and clear desk policies.	
7	Personal Devices	When personal devices are allowed for business use, enforce device configuration requirements and maintain control over data transmitted to and stored on the device.	
8	Intellectual Property	Protect material that can be considered intellectual property. Prevent copyright or licensing violations.	
9	Remote Work and Access	Apply additional restrictions and safeguards for remote work and remote access. Address the risks of working from, operating devices in, and transmitting data to off-premise locations.	
10	Termination and Continuity	Establish a process for the return of materials, de-provisioning of access, and handover of responsibilities upon termination of individual employment or a third-party contract.	
Physical Security (PS)			**AI Score**
1	Physical Access	Control physical access to facilities at defined access points and prevent unauthorized access through other points (e.g., windows, fire doors, delivery areas).	
2	Authorized Personnel	Ensure only authorized personnel, with proper identification, can access secure areas. Ensure visitors are escorted and their activity is monitored.	
3	Material Control	Inspect personal belongings and deliveries that are entering and leaving the facility.	
4	Environmental Threats	Identify and assess the physical environment, context, and related threats. Employ protections against hazards (fire, water, radiation, electromagnetic, tectonic, human activity).	
5	Utilities and Emergencies	Protect utilities such as power, gas, and telecommunications. Make emergency procedures readily available to personnel. Install emergency shutoffs and lighting.	
6	Monitoring and Alarms	Continuously monitor the physical premises and environmental conditions (temperature, humidity, etc.). Automatically respond or raise alarms when suspicious activity or abnormal conditions are detected.	

Privacy

The privacy category focuses on the organization's management of data, particularly personally identifiable information (PII), and practices to protect and control data throughout its life cycle.

Privacy Program (PP)			AI Score
1	Context	Understand the organization's legal and ethical obligations related to privacy, its role in the data processing ecosystem, and impacted stakeholders.	
2	Privacy Program	Establish a program to manage privacy risks, including that from third parties. Integrate the privacy program into the organization and measure its effectiveness.	
3	Personnel	Designate a privacy team to lead the implementation of the privacy program. Communicate to all personnel their roles and responsibilities with respect to privacy.	
4	Privacy-by-Design	Develop systems and practices based on security- and privacy-by-design principles.	
5	Notice and Consent	Be transparent about data and privacy practices. Publish a privacy policy in clear and understandable terms. Inform users when collecting their data and obtain consent.	
6	Data Minimization	Minimize the collection, processing, and use of personal data to what is absolutely necessary. Ensure personal data is only used for specified purposes.	
7	Privacy-enhancing Techniques	Employ privacy-enhancing techniques such as de-identification, anonymization, masking, encryption, differential privacy, and federated learning.	
8	Data Life cycle	Ensure the privacy of data across its life cycle (collection, authorization, processing, retention, and disposal). Map data actions and owners at each stage.	
9	Data Access and Separation	Enable granular access control and limit access to data. Segregate data that is mission critical, sensitive, or confidential.	
10	User Input and Control	Provide mechanisms for users to submit input, feedback, and grievances. Enable users to view, manage, and delete personal data collected about them.	

Handling PII (PI)			AI Score
1	Strategy and Oversight	Develop a strategy for handling PII based on the organization's role in the data processing ecosystem. Designate ownership over PII and the authorization of its processing. Implement and enforce privacy-by-design policies for handling PII.	
2	Purpose and Minimization	Define the purpose and legal basis for each PII processing activity. Limit collection and processing to what is strictly necessary. Securely dispose of PII when no longer needed.	
3	Notice and Transparency	Be transparent about practices regarding PII. Provide timely, concise, and easily accessible notice to individuals about the purpose of PII processing and details of how their PII will be handled at the time of collection. Make the information permanently accessible and regularly updated.	
4	Consent	Obtain explicit, informed consent from individuals before collecting and processing PII. Provide mechanisms to customize consent for specific purposes, update preferences, and withdraw consent.	
5	User Access and Control	Provide individuals with the ability to access, correct, request amendments to, and delete their PII retained by the organization. Enable individuals to object to PII processing and contest automated decisions made based on PII.	
6	Managing PII	Maintain accurate and up-to-date records of PII. Propagate corrections and deletions of PII data. Classify PII and use metadata tags to strictly track and control access to and use of PII. Retain secure backups of PII data.	
7	Assessing and Mitigating Risk	Conduct data and privacy impact assessments. Extend security controls to include privacy and the protection of PII to mitigate risks.	
8	Breaches and Notifications	Investigate security events where PII is involved to identify whether unauthorized access has occurred. Maintain a record of the investigation for auditability. When a breach occurs, notify impacted individuals and relevant authorities.	
9	Third-parties and Data Transfers	Maintain tight control over the authorized transfer and disclosure of PII to third parties. Document all transfers. Make a list of the possible third parties, countries, and international organizations that PII may be shared with available to individuals.	

10	Auditability and Compliance	Maintain the provenance of PII, the purpose for which it is used, authorizations, access, processing activities, transfer or disclosure, and disposal. Be able to demonstrate the compliance of these practices with applicable laws and regulations.	

Detection & Response

The detection and response section covers an organization's efforts to identify threats and incidents, respond when these events occur, and build greater operational continuity.

Audit Logging (LG)			**AI Score**
1	Audit Process and Scope	Define the scope of systems and events to be logged. Establish a process to generate, store, review, and analyze audit logs.	
2	What to Log	Log 1) access and modifications to data, software, and systems; 2) privileged actions; 3) other relevant personnel, user, and third-party activity; 4) system inputs and outputs; 5) errors; and 6) security events.	
3	Centralized Analysis	Integrate audit records across the organization into a centralized repository for analysis. Correlate information across multiple sources and monitoring activities.	
4	Monitor and Review Logs	Continuously review and monitor collected audit information to identify anomalous activity.	
5	Access and Security	Maintain the security and integrity of log data. Restrict access (read-only) to authorized personnel. Prevent log record modification and unauthorized disclosure or deletion.	
6	Storage and Capacity	Store log records separately from operational systems and ensure adequate storage capacity.	
7	Logging Failure	Alert personnel when audit logging mechanisms fail. Employ alternative logging capability, if available, or revert the system to a fail-safe mode (e.g., shutdown or limited functionality).	
Monitoring (MO)			**AI Score**
1	Process and Operations	Establish a security operations center responsible for monitoring and investigating security events. Adopt tools that facilitate the team's ability to collect information, prioritize analysis, and swiftly alert response teams.	
2	Threat Analysis	Perform threat analysis to identify the range of threat actors and their common attack vectors. Incorporate threat intelligence from information-sharing sources. Develop a monitoring strategy based on this analysis.	

3	Improving Defensive Posture	Preemptively strengthen the organization's cybersecurity posture by actively reducing the attack surface, employing predictive analytics, and establishing automated defenses.	0
4	Thresholds	Establish a security baseline for network activity, access, and system behavior to help identify anomalous activity. Continually review and revise monitoring thresholds and schedules.	24
5	Information Sharing	Share relevant threat intelligence, security events, and lessons learned during monitoring with internal and external stakeholders. Promote broader cybersecurity situational awareness.	35
6	Monitoring and Detection	Monitor physical and digital environments to detect anomalies, intrusions, security events, and potential insider threats. Employ deceptive techniques (e.g., honeypots) to detect intrusions and slow attackers.	21
7	Analyzing Event Data	Employ automated tools to support near-real-time analysis of event data. Centralize monitoring data (e.g., logs, reports, signatures, threat intel) for organization-wide visibility and analysis.	0
8	Alerts and Response	Establish an alert system to quickly notify relevant personnel when incidents are detected. Provide the monitoring team with the ability to rapidly lock down, restrict access, or take systems offline to prevent further compromise.	31
Incident Response (IR)			**AI Score**
1	Preparation	Create incident response and recovery plans. Identify stakeholders who will need to receive incident information. Designate responsibility for the execution of those plans. Conduct exercises to practice planned actions and assess their effectiveness.	38
2	Governance	Assign roles and responsibilities for incident response, including backups. Ensure the team is competently trained and has the requisite decision-making authority for response. Establish clear communication channels. Provide sufficient resourcing for response activities.	43
3	Detection and Identification	Define incident criteria and severity levels. Establish protected and confidential reporting mechanisms. Triage events and reported incidents. Initiate response actions when incident criteria has been met.	59
4	Response	Execute response plans upon detection of an incident: prioritize incidents, contain their impact, coordinate response with stakeholders, mitigate the cause, log response activities and evidence, and repair public relations.	26

5	Analysis and Investigation	During an incident, estimate the scope of its impact and identify its root cause. Log investigative actions and record evidence. After an incident, conduct a post-hoc assessment to identify trends and improve response effectiveness.	
6	Reporting	Report incident information to relevant authorities and stakeholders including affected communities, collaborative incident tracking initiatives, and information-sharing organizations.	
7	Recovery and Remediation	Coordinate recovery activities with internal and external stakeholders. Communicate progress on restoration and remediation. Verify satisfactory incident resolution before closure.	
8	Documentation and Logging	Maintain a repository of reported issues, near misses, incidents, and negative impacts. Document actions taken, outcomes, and performance metrics for response, recovery, and investigation activities.	
9	Lessons Learned	Collect and share lessons learned from incidents when they occur. Implement improvements to safety measures, security controls, and response plans based on post-hoc analysis and reviews.	
Resilience & Recovery (RR)			**AI Score**
1	Resilience and Continuity Program	Establish a business continuity strategy that prioritizes the critical function of the organization. Define resilience objectives and requirements.	
2	Resilience and Recovery Plans	Develop a business continuity and disaster recovery plan based on an analysis of potential threats, failures, and impacts. Regularly update the recovery plan and incorporate lessons learned.	
3	Dependencies and Third Parties	Identify the organization's essential function and its dependencies, particularly those related to suppliers. Coordinate and test response plans with those third parties.	
4	Capacity and Availability	Adequately maintain the organization's resource capacity. Identify constraints on capacity. Ensure capacity can be increased or demand decreased to maintain the availability of services.	
5	Backups	Create redundant copies of data and system configurations and store in a secure alternative location. Regularly test the backup process, the integrity of backups, and the restoration process.	
6	Resilience Mechanisms	Implement resilience mechanisms including redundancy (systems, services, equipment, etc.), fail-safes, failovers, load balancing, hot swapping, and alternative operating locations.	

7	Continuity of Utilities	Ensure the continuity of critical utilities (power, telecommunications, etc.). Consider using redundant services, backup sources, or alternative communication paths.	
8	Drills, Exercises, and Testing	Regularly test resilience mechanisms and recovery plans using automated testing, table-top exercises, realistic drills, and red teams. Ensure personnel have proficiency and proper training to conduct recovery activities.	
9	Restoration and Recovery Execution	Execute recovery plans to contain and mitigate events. Employ resilience mechanisms to ensure the continuity of critical functions. Restore systems and data, verifying their integrity.	

Insights

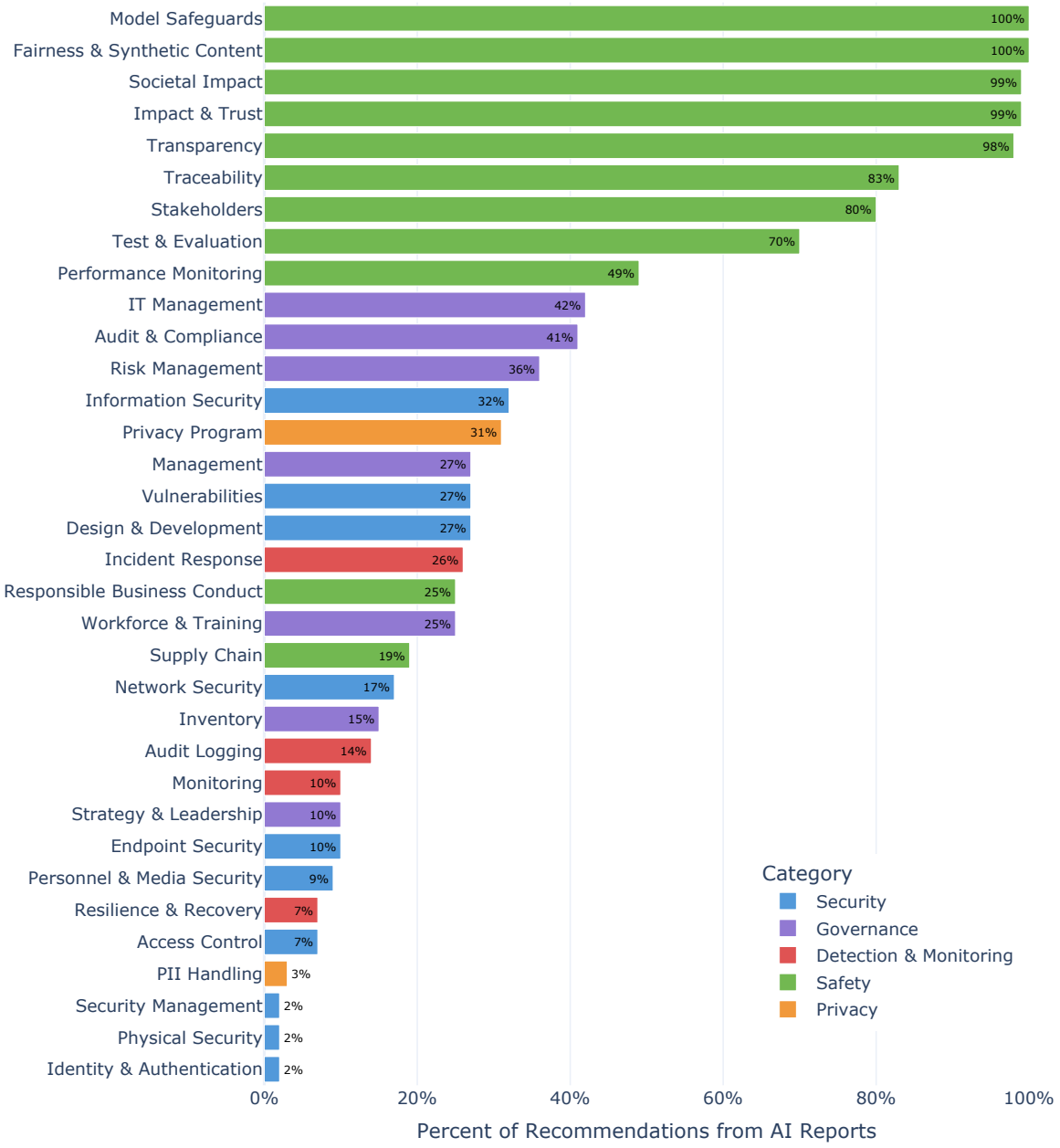
One of the key benefits of the harmonization process we developed to create this framework is the ability to trace the individual recommendations back to the set of original recommendations used to develop them. This information also enables us to draw several broader insights into how guidance from the 29 AI-specific reports compares to that of the 23 non-AI reports. In the following sections, we examine which topic areas are the focus of existing AI guidance documents and identify the gaps where further AI-specific guidance would be valuable.

The Focus of Existing AI Guidance Reports

The adoption of AI systems presents a new set of challenges for organizations to understand and manage. AI-related guidance documents, largely developed in the last few years, were created to help organizations address these new challenges. This raises the question, where does the guidance developed to date focus its attention?

Overall, we find that the majority of recommendations from AI reports (57.0%) pertain to Safety. Of the five high-level categories, Safety is the only one where the percent of recommendations from AI reports (81.8%) outweighs that from non-AI reports (18.2%). Comparatively, AI guidance comprises a minority of the recommendations in the Governance (29.9%), Security (15.4%), Detection & Response (15.3%), and Privacy (12.8%) categories. At the topic level, clusters within the Safety category tend to exhibit the highest proportion of recommendations from AI reports. However, we also observe concentrations of AI-related recommendations in several other areas outside of the Safety category. These include topics related to IT and risk management, compliance, information security, and the establishment of a privacy program. Figure 6 provides a full breakdown of the proportion of recommendations in each topic cluster that come from AI-specific reports. Each topic is color-coded by category, which further illustrates how AI-related guidance is heavily concentrated in the Safety category.

Figure 6: The Proportion of Recommendations Within Each Topic Cluster That Originate from a Report Specifically Related to AI



Source: CSET.

Building on overarching analysis of these results, we highlight five key areas where the attention of existing AI guidance has been focused:

1. **Expanded risks and impacts**: Most notable in existing guidance is the emphasis on the broader set of risks and impacts associated with AI systems. With the

large amount of data that is used in training these models, the probabilistic nature of their behavior, and the impact that these systems have—either directly or indirectly—on decision-making, it is no surprise that risk management must be a high priority for organizations adopting AI. Furthermore, with generative AI, these systems tend to be more user-facing, meaning that these risks can have greater direct impact on customers and stakeholders. This is in addition to substantial anticipated societal impacts, chemical, biological, radiological, and nuclear (CBRN) security concerns, and potential existential risks related to AI. The expanded risks and impacts can be observed in the Societal Impact, Impact & Trust, Risk Management, and Stakeholder topic areas.

2. **New vulnerabilities**: In addition to the vulnerabilities found in traditional software, AI systems introduce new attack vectors that adversaries can exploit. These systems are vulnerable to confidentiality attacks that extract information about the model (model theft or distillation) and the underlying training data (model inversion or membership inference). These systems can also be subject to integrity attacks that manipulate the behavior of the model to produce an adversary's desired output. These attacks include data poisoning, backdoors, adversarial input, and jailbreaking. Finally, adversaries can target these systems with availability attacks that use crafted inputs designed to make a model consume greater computational resources. While many cybersecurity and privacy principles help to address these vulnerabilities, new safeguards and techniques are needed. This is reflected in the proportion of AI-specific recommendations in the Model Safeguards, Information Security, and Privacy Program topic areas.
3. **Need for transparency**: Generally speaking, it is difficult to anticipate all of the possible outputs that a software system will produce.[19] The probabilistic nature of AI, the large input and output spaces of advanced models, and the purposeful inclusion of randomness in many generative AI tools in order to produce non-repetitive results further complicate this problem. In addition, many advanced models, such as neural networks and transformers, are considered to be black boxes in that the decisions that were reached cannot be explained in terms that humans can understand. Existing guidance highlights this gap and strongly recommends that organizations deploying AI systems in decision-making contexts provide mechanisms to provide impacted stakeholders with insight into the decision-making process and mechanisms to contest the output. This is evidenced by the Transparency & Oversight and Traceability topic areas.
4. **Greater testing and evaluation**: A heavy emphasis in existing guidance is placed on testing and evaluation capabilities for AI. This includes pre-deployment testing and continuous post-deployment monitoring and

evaluation. In particular, there is a focus on how benchmarking and new AI red-teaming techniques should be incorporated into existing organizational TEVV practices. These recommendations are found in the Test & Evaluation and Performance Monitoring topic areas.

5. **Synthetic content**: While computer-generated media, misinformation, and spam have been around for decades, generative AI has led to an explosion of synthetically generated text, images, audio, and video content.[20] Because these models can perpetuate underlying bias contained in training datasets, developers and deployers of these technologies must evaluate and closely monitor output for potential inappropriate, biased, or hateful content. However, regardless of whether an organization chooses to adopt this technology, it will be faced with the challenge of differentiating synthetic and real content. Organizations and individuals must be aware of the use of synthetic content for manipulation and deception, enhancing threats from misinformation, deepfakes, and social engineering attacks. Recommendations pertaining to synthetic content are primarily located in the Fairness & Synthetic Content topic area.

In part, the concentration of recommendations in these topics can be interpreted as areas where AI guidance is more established, readily available, or easily adoptable. Alternatively, these results may also provide an indication of where experts believe the most substantial AI challenges exist and therefore where organizations should be focusing their attention. In reality, it is likely a combination of these factors. While these hypotheses are speculative, existing AI guidance has clearly drawn heavily from the AI safety and AI trust communities. AI security is not absent from existing guidance—one of the five focus areas discussed above relates to the novel vulnerabilities in AI systems—yet the large imbalance in attention suggests that it would be worth revisiting whether further work may be needed on issues pertaining to AI security and, if so, what barriers have prevented such guidance from being developed.

Where There Are Gaps in AI Guidance

Beyond the notable imbalance between AI safety and security, there are several additional topics that the research team felt were missing from existing AI guidance. Our analysis is not exhaustive, so existing guidance relative to these areas may exist elsewhere. However, the relative absence of guidance on these topics in the reports we examined is nonetheless concerning given the central role they have played in policy discussions. These topic areas include:

1. **Workforce**: While existing guidance discusses the societal impacts that AI will have on the workforce, there are little to no recommendations related to how organizations should be addressing that impact internally. Organizations need to be thinking about how worker displacement may affect their own employees and develop strategies to manage these changes, such as through re-skilling initiatives. Furthermore, there is a lack of guidance on how to upskill an organization's existing workforce to competently use and manage AI tools. There is a similar gap related to awareness training for employees that covers synthetic content and related risks, responsible use of AI tools, and general AI literacy. Such guidance would relate directly to the Workforce & Training topic area.
2. **Incident reporting**: Transparency and communication related to AI incidents has been a central topic of discussion among policymakers and the AI safety community. Incident reporting mechanisms exist in many industries. Some of these forums are voluntary, while others are legally required depending on sector, jurisdiction, and type of incident (e.g., safety, cybersecurity, privacy). However, reporting requirements can become murky when incidents involve AI. Some AI-related incidents plainly fall under the umbrella of a cybersecurity incident, safety violation, or privacy breach and should be handled as such. Others cases are not so clear. While organizations should leverage existing structures and internal incident management teams, these structures may need to be updated or expanded to account for AI-related incidents. Further guidance on how to best capture AI incidents through available reporting mechanisms and how to handle AI incidents that may not neatly fall into existing buckets would be valuable to organizations. This information would be relevant to the Incident Response topic.
3. **Confidential and privacy-sensitive information**: The leaking of confidential, proprietary, and privacy-sensitive information through the use of chatbots and other AI-enabled tools is a serious concern for organizations. Yet, while there is a substantial amount of guidance that covers how personnel should protect this information during in-person conversations, telephone communications, email, and even fax, there are no corresponding recommendations for managing risks through interactions with AI systems. This guidance would pertain to the Personnel & Media Security and Handling PII topics.
4. **Agentic AI**: Existing AI-specific guidance almost wholly pertains to LLMs and generative AI. However, with the rapid pace of AI development, organizations need to be forward thinking and therefore guidance must be as well. The automation of workflows using AI agents—AI that can plan and take action in the real world—is likely on the near horizon.[21] For these agents to be useful,

they will need to be able to access a variety systems and assets belonging to the organization. Managing that access, maintaining identities for various agents, and tracking their actions across the organization will be critical. This information would be relevant to the Audit Logging, Access Control, and Identity & Authentication topics areas, among others.

Conclusion

In this report, we present a harmonized set of recommended practices based on the analysis of 52 existing frameworks on artificial intelligence, safety, cybersecurity, privacy, and risk management. This framework represents a distillation of the collective knowledge of 7,741 recommended practices, covering a much broader scope than any existing report individually. Our set of 258 harmonized recommendations provides organizations a single resource for adopting a comprehensive approach to the management of technology and the adoption of AI systems. These recommendations are neatly organized into 34 topic areas and grouped into five overarching categories, enabling organizations to easily identify and prioritize the most important practices relevant to their use case. In developing this resource, we provide and validate a mixed-methods approach to harmonization that can be reused and applied to other domains. Based on the harmonization results, we provide insight into the areas in which existing AI guidance is concentrated and where there are gaps.

This report represents the first step toward addressing the challenges organizations face in implementing AI guidance. In synthesizing a single, clearly written, relatively small yet comprehensive set of recommendations from existing guidance, we help to address challenges related to information overload, disparate sources of information, and inaccessible language. Moving forward, this framework will serve as the foundation for future CSET work aimed at providing AI-specific implementation details and tailoring that guidance to several use cases.

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.

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.

Ali Crawford is a senior research analyst at CSET, working on the CyberAI Project focusing on how the United States is building and maintaining cyber and AI education and workforce ecosystems.

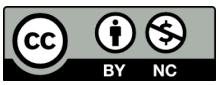
Mia Hoffmann is a research fellow at CSET, where her work focuses on AI governance.

Colin Shea-Blymyer is a research fellow at CSET, where he works on the CyberAI Project focusing on AI red-teaming.

John Bansemer is a non-resident senior fellow at CSET and is also an adjunct professor at Georgetown University's School of Foreign Service.

Acknowledgements

We would like to thank Catherine Aiken, Matt Mahoney, Jessica Newman, 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: 10.51593/20240041

Appendix

List of Guidance Documents Examined

Table 1: List of Examined Guidance Documents and Publishing Organizations

Organization	Crossref	Report
AICPA	AICPA	2017 Trust Services Criteria for Security, Availability, Processing Integrity, Confidentiality, and Privacy[22]
AI Verify	AI VERIFY	AI Governance Testing Framework and Toolkit[23]
Center for Internet Security (CIS)	CIS	CIS Critical Security Controls[24]
Cloud Security Alliance (CSA)	CCM	Cloud Controls Matrix v4.0.12[25]
Cybersecurity and Infrastructure Security Agency (CISA)	CISA CPG	CPG: Cross-Sector Cybersecurity Performance Goals[26]
Cybersecurity and Infrastructure Security Agency (CISA)	CISA CR	Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Security-by-Design and -Default[27]
Cyber Risk Institute	CRI	The CRI Profile Version 2.0[28]
Department of Homeland Security (DHS)	DHS	Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure[29]
Department of State (DoS)	DOS	Risk Management Profile for Artificial Intelligence and Human Rights[30]
European Union Agency for Cybersecurity (ENISA)	ENISA	Securing Machine Learning Algorithms[31]
European Commission	EU TAI	Ethics Guidelines for Trustworthy AI[32]
Future of Life Institute	FLI	Asilomar AI Principles[33]
Google	SAIF	Google's Secure AI Framework[34]
Hiroshima AI Process	ICCAI	Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems[35]
International Association of Privacy Professionals (IAPP)	IAPP	Certified Information Privacy Manager Body of Knowledge and Exam Blueprint Version 4.1[36]

International Organization for Standardization (ISO)	ISO 23894	ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management[37]
International Organization for Standardization (ISO)	ISO 27001	ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection — Information security management systems — Requirements[38]
International Organization for Standardization (ISO)	ISO 27002	ISO/IEC 27002:2022 Information security, cybersecurity and privacy protection — Information security controls[39]
International Organization for Standardization (ISO)	ISO 27701	ISO/IEC 27701:2019 Security techniques — Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management — Requirements and guidelines[40]
International Organization for Standardization (ISO)	ISO 31000	ISO/IEC 31000:2018 Risk management — Guidelines[41]
International Organization for Standardization (ISO)	ISO 42001	ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system[42]
ISACA	COBIT 5	COBIT 5: A Business Framework for the Governance and Management of Enterprise IT[43]
Japan's Ministry of Internal Affairs and Communication; Ministry of Economy, Trade, and Industry	AI GFB	AI Guidelines for Business Version 1.0[44]
National Institute of Standards and Technology (NIST)	NIST AI PB	AI RMF Playbook[45]
National Institute of Standards and Technology (NIST)	NIST AI RMF	Artificial Intelligence Risk Management Framework (AI RMF 1.0)[46]
National Institute of Standards and Technology (NIST)	NIST GAI	Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile[47]
National Institute of Standards and Technology (NIST)	NIST CIC	Framework for Improving Critical Infrastructure Cybersecurity Version 1.1[48]
National Institute of Standards and Technology (NIST)	NIST ICT	Information and Communications Technology (ICT) Risk Outcomes[49]
National Institute of Standards and Technology (NIST)	NIST CSF	The NIST Cybersecurity Framework (CSF) 2.0[50]

National Institute of Standards and Technology (NIST)	NIST PF	NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management Version 1.0[51]
National Institute of Standards and Technology (NIST)	NIST RMF	Risk Management Framework for Information Systems and Organizations Revision 2[52]
National Institute of Standards and Technology (NIST)	NIST SSDF GAI	Secure Software Development Practices for Generative AI and Dual-Use Foundation Models[53]
National Institute of Standards and Technology (NIST)	NIST SSDF	Secure Software Development Framework (SSDF) Version 1.1[54]
National Institute of Standards and Technology (NIST)	NIST SP 800-53	Security and Privacy Controls for Information Systems and Organizations Revision 5[55]
National Security Agency's Artificial Intelligence Security Center (NSA AISC)	AI DPLY	Deploying AI Systems Securely[56]
The Organization for Economic Co-operation and Development (OECD)	OECD DDG	OECD Due Diligence Guidance for Responsible Business Conduct[57]
The Organization for Economic Co-operation and Development (OECD)	OECD AIP	OECD AI Principles[58]
Office of Management and Budget (OMB)	OMB A-130	Circular A-130 Managing Information as a Strategic Resource [59]
Open Worldwide Application Security Project (OWASP)	OWASP ML	OWASP Machine Learning Security Top Ten Version 0.3[60]
Open Worldwide Application Security Project (OWASP)	OWASP LLM	OWASP Top 10 for LLM Applications Version 1.1[61]
Partnership on AI (PAI)	PAI SFMD	PAI's Guidance for Safe Foundation Model Deployment [62]
Partnership on AI (PAI)	PAI RPSM	PAI's Responsible Practices for Synthetic Media[63]
Responsible AI Institute (RAII)	RAII	Best Practices in Generative AI Responsible use and development in the modern workplace[64]
Singapore Personal Data Protection Commission (PDPC)	PDPC	Model Artificial Intelligence Governance Framework Second Edition[65]
The Software Alliance (BSA)	BSA	Confronting Bias: BSA's Framework to Build Trust

		in AI[66]
Center for Long-term Cybersecurity	TTAI	A Taxonomy of Trustworthiness for Artificial Intelligence[67]
U.K. National Cyber Security Centre (NCSC)	NCSC CAF	Cyber Assessment Framework Version 3.2[68]
U.K. National Cyber Security Centre (NCSC); U.S. Cybersecurity and Infrastructure Security Agency (CISA)	AI DEV	Guidelines for secure AI system development[69]
U.K. National Cyber Security Centre (NCSC)	NCSC ML	Principles for the security of machine learning[70]
U.K. National Cyber Security Centre (NCSC)	NCSC SC	Supply chain security guidance[71]
United Nations Educational, Scientific and Cultural Organization (UNESCO)	UNESCO AI	Recommendation on the Ethics of Artificial Intelligence[72]
University of Turku	TURKU	Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance[73]

Example of Standardization

Below is an example of the standardization process applied to an individual recommendation sampled from the DHS report *Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure*.[74] The original text of the recommendation and the final standardized text used for clustering are shown below.

Original: Ensure alignment with human-centric values. AI model developers should ensure, to the best of their ability, that AI models reflect human values and goals, with the ultimate objective of ensuring they are helpful, accurate, unbiased, and transparent.[18] AI application developers should align use cases with values that respect civil rights, civil liberties, and applicable laws in partnership with relevant civil society.[19]

Standardized: Ensure alignment with human-centric values. Ensure, to the best of the organization's ability, that systems reflect human values and goals, with the ultimate objective of ensuring they are helpful, accurate, unbiased, and transparent. Align use cases with values that respect civil rights, civil liberties, and applicable laws in partnership with relevant civil society

To transform the original into the standardized version the following steps were taken:

1. Removing references and placeholder text

   Here we remove the references to external documents included in the recommendation.

   Ensure alignment with human-centric values: AI model developers should ensure, to the best of their ability, that AI models reflect human values and goals, with the ultimate objective of ensuring they are helpful, accurate, unbiased, and transparent. AI application developers should align use cases with values that respect civil rights, civil liberties, and applicable laws in partnership with relevant civil society.

2. Standardizing the audience

   Here we standardize the audience for the recommendation by replacing "AI model developers" and "AI application developers" with "the organization."

   Ensure alignment with human-centric values. **The organization** should ensure, to the best of its ability, that AI models reflect human values and goals, with the

ultimate objective of ensuring they are helpful, accurate, unbiased, and transparent. **The organization** should align use cases with values that respect civil rights, civil liberties, and applicable laws in partnership with relevant civil society.

3. Masking “artificial intelligence”

   Here we replace references to artificial intelligence. In this example, we replace the term “AI models” with “systems.”

   Ensure alignment with human-centric values. The organization should ensure, to the best of its ability, that systems reflect human values and goals, with the ultimate objective of ensuring they are helpful, accurate, unbiased, and transparent. The organization should align use cases with values that respect civil rights, civil liberties, and applicable laws in partnership with relevant civil society.

4. Converting to the active voice

   Here we convert the passive voice to the active voice in the second and third lines.

   Ensure alignment with human-centric values. Ensure, to the best of the organization’s ability, that systems reflect human values and goals, with the ultimate objective of ensuring they are helpful, accurate, unbiased, and transparent. Align use cases with values that respect civil rights, civil liberties, and applicable laws in partnership with relevant civil society.

Summary of Clustering and Harmonization Results

Table 2: Summary of Clustering and Harmonization Results by Topic

Cluster	Contributing Reports	Original Recommendations	Percent of Corpus	Harmonized Recommendations
Design & Development	23	485	6.3%	10
Risk Management	31	444	5.7%	10
Incident Response	22	326	4.2%	9
Personnel & Media Security	14	323	4.2%	10
Resilience & Recovery	39	316	4.1%	9
PII Handling	23	315	4.1%	10
Information Security	11	307	4.0%	9
Traceability	16	284	3.7%	6
Supply Chain	24	278	3.6%	7
Management	36	274	3.5%	8
Transparency & Oversight	36	272	3.5%	7
Performance Monitoring	24	266	3.4%	5
Network Security	29	243	3.1%	8
Impact & Trust	21	242	3.1%	8
IT Management	22	239	3.1%	10
Workforce & Training	22	239	3.1%	5
Endpoint Security	28	215	2.8%	8
Stakeholders	37	206	2.7%	6
Model Safeguards	26	205	2.6%	6
Monitoring	10	191	2.5%	8
Identity & Authentication	14	190	2.5%	9
Security Management	28	190	2.5%	8
Access Control	27	187	2.4%	9
Privacy	7	183	2.4%	10
Audit & Compliance	37	178	2.3%	7
Test & Evaluation	18	162	2.1%	8
Societal Impact	17	158	2.0%	7
Audit Logging	26	147	1.9%	7
Inventory	31	126	1.6%	5
Bias, Fairness, & Synthetic Content	28	125	1.6%	5
Governance	32	110	1.4%	5
Responsible Business Conduct	28	110	1.4%	6
Vulnerabilities	23	108	1.4%	7
Physical Security	35	97	1.3%	6

Table 3: Summary of Clustering and Harmonization Results by Category

Cluster	Contributing Reports	Original Recommendations	Percent of Corpus	Harmonized Recommendations
Security	42	2,345	30.3%	84
Safety	49	2,030	26.2%	58
Governance	46	1,888	24.4%	63
Detection & Response	42	980	12.7%	33
Privacy	29	498	6.4%	20

Endnotes

[1] Anna Jobin, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence* 1 (2019): 389–399, https://doi.org/10.1038/s42256-019-0088-2.

[2] Exec. Order No. 14110, 88 FR 75191 (2023), rescinded 20 January 2025, https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

[3] 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; 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.

[4] Kevin Poireault, "UK's AI Safety Institute Rebrands amid Government Strategy Shift," *Infosecurity Magazine*, February 14, 2025, https://www.infosecurity-magazine.com/news/uk-ai-safety-institute-rebrands/; Madison Adler, "Trump Administration Rebrands AI Safety Institute," *Fedscoop*, June 4, 2025, https://fedscoop.com/trump-administration-rebrands-ai-safety-institute-aisi-caisi/.

[5] Xiangyu Qi, Yi Zeng, Edoardo Debenedetti et al., "AI Risk Management Should Incorporate Both Safety and Security," arXiv preprint arXiv:2405.19524 (May 2024), https://arxiv.org/abs/2405.19524.

[6] J.D. Vance, "Remarks by the Vice President at the Artificial Intelligence Action Summit in Paris, France," The American Presidency Project, February 11, 2025, https://www.presidency.ucsb.edu/documents/remarks-the-vice-president-the-artificial-intelligence-action-summit-paris-france.

[7] Exec. Order No. 14179, 90 FR 8741 (2025), https://www.federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence.

[8] Office of Management and Budget, *M-25-21: Accelerating Federal Use of AI through Innovation, Governance, and Public Trust* (Washington, DC: Executive Office of the President, April 3, 2025), https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf; Office of Management and Budget, *M-25-22: Driving Efficient Acquisition of Artificial Intelligence in Government* (Washington, DC: Executive Office of the President, April 3, 2025), https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-22-Driving-Efficient-Acquisition-of-Artificial-Intelligence-in-Government.pdf; Office of Management and Budget, *M-24-10: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence* (Washington, DC: Executive Office of the President, March 28, 2024), https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf; Office of Management and Budget, *M-24-18: Advancing the Responsible Acquisition of Artificial Intelligence in Government*

(Washington, DC: Executive Office of the President, September 24, 2024), https://www.whitehouse.gov/wp-content/uploads/2024/10/M-24-18-AI-Acquisition-Memorandum.pdf.

[9] NIST, *Artificial Intelligence Risk Management Framework* (Washington, DC: Department of Commerce, 2023), https://www.nist.gov/itl/ai-risk-management-framework; NIST, *Cybersecurity Framework 2.0* (Washington, DC: Department of Commerce, 2024), https://www.nist.gov/cyberframework; NIST, *Privacy Framework* (Washington, DC: Department of Commerce, 2020), https://www.nist.gov/privacy-framework.

[10] NIST, *AI RMF Playbook* (Washington, DC: Department of Commerce, 2024), https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook.

[11] NIST, *CSF 2.0 Profiles* (Washington, DC: Department of Commerce, 2025), https://www.nist.gov/cyberframework/profiles; NIST, *AI RMF Profiles* (Washington, DC: Department of Commerce, 2024), https://www.nist.gov/itl/ai-risk-management-framework/ai-risk-management-framework-resources.

[12] Chief Digital and Artificial Intelligence Office, *RAI Toolkit* (Washington, DC: Department of Defense, 2025), https://rai.tradewindai.com/assessment; "PAI's Guidance for Safe Foundation Model Deployment," Partnership on AI, 2024, https://partnershiponai.org/modeldeployment/.

[13] Kyle Crichton, Jessica Ji, Kyle Miller et al., "Securing Critical Infrastructure in the Age of AI" (Center for Security and Emerging Technology, October 2024), https://cset.georgetown.edu/publication/securing-critical-infrastructure-in-the-age-of-ai/.

[14] "Vector embeddings" OpenAI, 2024, https://platform.openai.com/docs/guides/embeddings.

[15] Alina Petukhova, João P. Matos-Carvalho, and Nuno Fachada, "Text Clustering with Large Language Model Embeddings," *International Journal of Cognitive Computing in Engineering* 6 (2025): 100–108, https://doi.org/10.1016/j.ijcce.2024.11.004.

[16] Kathy Charmaz, "Grounded Theory as an Emergent Method," in *Handbook of Emergent Methods* (New York: Guilford Press, 2008), 155-172, http://www.sxf.uevora.pt/wp-content/uploads/2013/03/Charmaz_2008-b.pdf.

[17] Cliodhna O'Connor and Helene Joffe, "Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines," *International Journal of Qualitative Methods* 19 (2020) https://doi.org/10.1177/1609406919899220; Andrei P. Kirilenko and Svetlana Stepchenkova, "Inter-Coder Agreement in One-to-Many Classification: Fuzzy Kappa," *PLoS ONE* 11, no. 3 (March 2016), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0149787.

[18] Matthew B. Miles and A. Michael Huberman, *Qualitative Data Analysis: An Expanded Sourcebook* (Thousand Oaks, CA: SAGE Publications, 1994).

[19] D.C. Kozen, "Rice's Theorem," in *Automata and Computability* (New York: Springer Science, 1977), https://doi.org/10.1007/978-3-642-85706-5_42.

[20] Finn Brunton, *Spam: A Shadow History of the Internet* (Cambridge, MA: The MIT Press, 2013), https://doi.org/10.7551/mitpress/9384.001.0001.

[21] Helen Toner, John Bansemer, Kyle Crichton et al., "Through the Chat Window and into the Real World: Preparing for AI Agents" (Center for Security and Emerging Technology, October 2024), https://cset.georgetown.edu/publication/through-the-chat-window-and-into-the-real-world-preparing-for-ai-agents/.

[22] AICPA, *2017 Trust Services Criteria for Security, Availability, Processing Integrity, Confidentiality, and Privacy (with Revised Points of Focus – 2022)* (AICPA, 2022), https://www.aicpa-cima.com/resources/download/2017-trust-services-criteria-with-revised-points-of-focus-2022.

[23] AI Verify, "AI Governance Testing Framework and Toolkit," *Summary Report Binary Classification Model for Credit Risk ABC Company PTE LTD* (AI Verify, 6 June 2023), https://aiverifyfoundation.sg/downloads/AI_Verify_Sample_Report.pdf.

[24] Center for Internet Security, *CIS Critical Security Controls Version 8.1* (CIS, August 2024), https://www.cisecurity.org/controls/v8-1.

[25] Cloud Security Alliance, *CSA Cloud Controls Matrix v4.0.12* (CSA, 3 June 2024), https://cloudsecurityalliance.org/research/cloud-controls-matrix.

[26] Cybersecurity and Infrastructure Security Agency, *CPG: Cross-Sector Cybersecurity Performance Goals* (Washington, DC: CISA, March 2023), https://www.cisa.gov/sites/default/files/2023-03/CISA_CPG_REPORT_v1.0.1_FINAL.pdf.

[27] Cybersecurity and Infrastructure Security Agency, *Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Security-by-Design and -Default* (Washington, DC: CISA, 13 April 2023), https://www.cisa.gov/sites/default/files/2023-04/principles_approaches_for_security-by-design-default_508_0.pdf.

[28] The Cyber Risk Institute, *CRI Profile Version 2.0* (CRI, 29 February 2024), https://cyberriskinstitute.org/the-profile/.

[29] U.S. Department of Homeland Security and the Artificial Intelligence Safety and Security Board, *Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure* (Washington, DC: Department of Homeland Security, November 14, 2024), https://www.dhs.gov/sites/default/files/2024-11/24_1114_dhs_ai-roles-and-responsibilities-framework-508.pdf.

[30] U.S. Department of State, *Risk Management Profile for Artificial Intelligence and Human Rights* (Washington, DC: Department of State, 25 July 2024), https://2021-2025.state.gov/risk-management-profile-for-ai-and-human-rights/.

[31] European Union Agency for Cybersecurity, *Securing Machine Learning Algorithms* (ENISA, December 2021), https://doi.org/10.2824/874249.

[32] Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission, *Ethics Guidelines for Trustworthy AI* (8 April 2019), https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

[33] Future of Life Institute, "Asilomar AI Principles" (FLI, 11 August 2017), https://futureoflife.org/open-letter/ai-principles/.

[34] Google, "Google's Secure AI Framework" (Google, June 2023), https://saif.google/.

[35] Hiroshima AI Process, *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems* (Hiroshima, Japan: October 2023), https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document05_en.pdf.

[36] International Association of Privacy Professionals, *Certified Information Privacy Manager Body of Knowledge and Exam Blueprint Version 4.1* (IAPP, 2 September 2024), https://iapp.org/media/pdf/certification/CIPM_BoK_EBP_%204.1.0_FINAL_PROOFED_CLEAN.pdf.

[37] International Organization for Standardization, *ISO/IEC 23894:2023 Information Technology — Artificial Intelligence — Guidance on Risk Management* (ISO, February 2023), https://www.iso.org/standard/77304.html.

[38] International Organization for Standardization, *ISO/IEC 27001:2022 Information Security, Cybersecurity and Privacy Protection — Information Security Management Systems — Requirements* (ISO, October 2022), https://www.iso.org/standard/27001.

[39] International Organization for Standardization, *ISO/IEC 27002:2022 Information Security, Cybersecurity and Privacy Protection — Information Security Controls* (ISO, March 2022), https://www.iso.org/standard/75652.html.

[40] International Organization for Standardization, *ISO/IEC 27701:2019 Security Techniques — Extension to ISO/IEC 27001 and ISO/IEC 27002 for Privacy Information Management — Requirements and Guidelines* (ISO, August 2019), https://www.iso.org/standard/71670.html.

[41] International Organization for Standardization, *ISO/IEC 31000:2018 Risk Management — Guidelines* (ISO, February 2018), https://www.iso.org/standard/65694.html.

[42] International Organization for Standardization, *ISO/IEC 42001:2023 Information Technology — Artificial Intelligence — Management System* (ISO, December 2023), https://www.iso.org/standard/42001.

[43] ISACA, *COBIT 5: A Business Framework for the Governance and Management of Enterprise IT* (ISACA, 2012), https://www.isaca.org/resources/cobit/cobit-5.

[44] Japan's Ministry of Internal Affairs and Communication and Ministry of Economy, Trade, and Industry, *AI Guidelines for Business Ver1.0* (19 April 2024), https://www.soumu.go.jp/main_content/000943087.pdf.

[45] NIST, *AI RMF Playbook* (Washington, DC: Department of Commerce, 2024), https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook.

[46] NIST, *Artificial Intelligence Risk Management Framework* (Washington, DC: Department of Commerce, 2023), https://www.nist.gov/itl/ai-risk-management-framework.

[47] NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (Washington, DC: Department of Commerce, July 2024), https://doi.org/10.6028/NIST.AI.600-1.

[48] NIST, Framework for Improving Critical Infrastructure Cybersecurity Version 1.1 (NIST, April 2018), https://www.nist.gov/cyberframework/csf-11-archive.

[49] Stephen Quinn, Nahla Ivy, Julie Chua et al., *Information and Communications Technology (ICT) Risk Outcomes: Integrating ICT Risk Management Programs with the Enterprise Risk Portfolio* (Washington, DC: NIST, Department of Commerce, November 2023), https://doi.org/10.6028/NIST.SP.800-221A.

[50] NIST, *Cybersecurity Framework 2.0* (Washington, DC: Department of Commerce, 2024), https://www.nist.gov/cyberframework.

[51] NIST, *Privacy Framework* (Washington, DC: Department of Commerce, 2020), https://www.nist.gov/privacy-framework.

[52] NIST, *Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy* (Washington, DC: Department of Commerce, December 2018), Revision 2, https://doi.org/10.6028/NIST.SP.800-37r2.

[53] Harold Booth, Murugiah Souppaya, Apostol Vassilev et al., *Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile* (Washington, DC: NIST, Department of Commerce, July 2024), https://doi.org/10.6028/NIST.SP.800-218A.

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

[55] NIST, *Security and Privacy Controls for Information Systems and Organizations* (Washington, DC: Department of Commerce, September 2020), Revision 5, https://doi.org/10.6028/NIST.SP.800-53r5.

[56] National Security Agency's Artificial Intelligence Security Center, *Deploying AI Systems Securely* (NSA AISC, 15 April 2024), https://media.defense.gov/2024/Apr/15/2003439257/-1/-1/0/CSI-DEPLOYING-AI-SYSTEMS-SECURELY.PDF.

[57] OECD, *OECD Due Diligence Guidance for Responsible Business Conduct* (OECD, 2018), https://mneguidelines.oecd.org/OECD-Due-Diligence-Guidance-for-Responsible-Business-Conduct.pdf.

[58] OECD, "OECD AI Principles Overview" (OECD, May 2024), https://oecd.ai/en/ai-principles.

[59] White House Office of Management and Budget, "Managing Information as a Strategic Resource" (Washington, DC: OMB, December 1985), Circular A-310, https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/circulars/A130/a130revised.pdf.

[60] The Open Worldwide Application Security Project, "OWASP Machine Learning Security Top Ten" (OWASP, 2023), Version 0.3, https://owasp.org/www-project-machine-learning-security-top-10/.

[61] The Open Worldwide Application Security Project, *OWASP Top 10 for LLM Applications* (OWASP, 16 October 2023), Version 1.1, https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_1.pdf.

[62] Partnership on AI, "PAI's Guidance for Safe Foundation Model Deployment: A Framework for Collective Action" (PAI, July 2024), https://partnershiponai.org/modeldeployment/.

[63] Partnership on AI, *PAI's Responsible Practices for Synthetic Media* (PAI, 27 February 2023), https://partnershiponai.org/wp-content/uploads/2023/02/PAI_synthetic_media_framework.pdf.

[64] Responsible Artificial Intelligence Institute, *Best Practices in Generative AI Responsible Use and Development in the Modern Workplace* (RAII, 2024), https://www.responsible.ai/best-practices-in-generative-ai-guide/.

[65] Singapore Personal Data Protection Commission, *Model Artificial Intelligence Governance Framework Second Edition* (Singapore: PGPC, January 2020), https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf.

[66] The Software Alliance, *Confronting Bias: BSA's Framework to Build Trust in AI* (BSA, 8 June 2021), https://ai.bsa.org/confronting-bias-bsas-framework-to-build-trust-in-ai/.

[67] Jessica Newman, *A Taxonomy of Trustworthiness for Artificial Intelligence* (Center for Long-term Cybersecurity, UC Berkely, January 2023), https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence/.

[68] U.K. National Cyber Security Centre, *Cyber Assessment Framework V3.2* (NCSC, 15 April 2024), https://www.ncsc.gov.uk/collection/cyber-assessment-framework.

[69] U.K. National Cyber Security Centre and U.S. Cybersecurity and Infrastructure Security Agency, *Guidelines for Secure AI System Development* (NCSC and CISA, 2023), https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf.

[70] U.K. National Cyber Security Centre, *Principles for the Security of Machine Learning* (NCSC, August 2022), https://www.ncsc.gov.uk/files/Principles-for-the-security-of-machine-learning.pdf.

[71] U.K. National Cyber Security Centre, "Supply Chain Security Guidance" (NCSC, 28 January 2018), https://www.ncsc.gov.uk/collection/supply-chain-security.

[72] The United Nations Educational, Scientific and Cultural Organization, *Recommendation on the Ethics of Artificial Intelligence* (UNESCO, 23 November 2021), https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence.

[73] Matti Mäntymäki, Matti Minkkinen, Teemu Birkstedt, and Mika Viljanen, "Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance" (ArXiv, February 2023), https://doi.org/10.48550/arXiv.2206.00335.

[74] U.S. Department of Homeland Security and the Artificial Intelligence Safety and Security Board, *Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure* (Washington, DC: Department of Homeland Security, November 14, 2024), https://www.dhs.gov/sites/default/files/2024-11/24_1114_dhs_ai-roles-and-responsibilities-framework-508.pdf.