Artificial Intelligence (AI) Policy

[Agency]

I. Purpose

This policy establishes a comprehensive, yet flexible, governance structure for AI systems used by, or on behalf of, the [Agency]. This policy enables the [Agency] to use AI systems for the benefit of the community while safeguarding against potential harms.

The key objectives of the AI Policy are to:

- Provide guidance that is clear, easy to follow, and supports decision-making for the staff (full-time, part-time), interns, consultants, contractors, partners, and volunteers who may be purchasing, configuring, developing, operating, or maintaining the [Agency's] AI systems or leveraging AI systems to provide services to the [Agency].
- Ensure that when using AI systems, the [Agency] or those operating on its behalf, adhere to the Guiding Principles that represent values with regards to how AI systems are purchased, configured, developed, operated, or maintained.
- Define roles and responsibilities related to the [Agency's] usage of AI systems.
- Establish and maintain processes to assess and manage risks and values presented by AI systems used by the [Agency];
- Align the governance of AI systems with existing data governance, security, and privacy measures in accordance with the [Agency's Information Security Policy] and [Agency's Data Governance Policy].
- Define prohibited uses of AI systems;
- Establish "sunset" procedures to safely retire AI systems that no longer meet the needs of the [Agency];
- Define how AI systems may be used for legitimate [Agency] purposes in accordance with applicable local, state, and federal laws, and existing agency policies.

The [Agency] defines "artificial intelligence" or "AI" to be a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.[1] AI systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action.

ATTACHMENT 2

[1] Definition from 15 U.S.C. 9401(3).

The [Agency] defines an "AI system" to be any data system, software, hardware, application, tool, or utility that operates in whole or in part using AI.

The [Agency's] AI systems and the data contained therein will be purchased, configured, developed, operated, and maintained using the [Agency's] [AI Governance Manual, or applicable handbook].

II. Scope

This policy applies to:

1. All AI systems deployed by the [Agency] or on the [Agency's] behalf; and

2. Staff (full-time, part-time), interns, consultants, contractors, partners, and volunteers who may be purchasing, configuring, developing, operating, or maintaining the [Agency's] AI systems or who may be leveraging AI systems to provide services to the [Agency].

III. Guiding Principles for Responsible AI Systems

These principles describe the [Agency's] values with regards to how AI systems are purchased, configured, developed, operated, or maintained.

1. **Human-Centered Design:** AI systems are developed and deployed with a human-centered approach that evaluates AI powered services for their impact on the public.

2. **Security & Safety:** AI systems maintain confidentiality, integrity, and availability through safeguards that prevent unauthorized access and use. Implementation of AI systems is reliable and safe, and minimizes risks to individuals, society, and the environment.

3. **Privacy:** Privacy is preserved in all AI systems by safeguarding personally identifiable information (PII) and sensitive data from unauthorized access, disclosure, and manipulation.

4. **Transparency:** The purpose and use of AI systems is proactively communicated and disclosed to the public. An AI system, its data sources, operational model, and policies that govern its use are understandable and documented.

5. **Equity:** AI systems support equitable outcomes for everyone. Bias in AI systems is effectively managed with the intention of reducing harm for anyone impacted by its use.

6. **Accountability:** Roles and responsibilities govern the deployment and maintenance of AI systems, and human oversight ensures adherence to relevant laws and regulations.

7. **Effectiveness:** AI systems are reliable, meet their objectives, and deliver precise and dependable outcomes for the utility and contexts in which they are deployed.

8. **Workforce Empowerment:** Staff are empowered to use AI in their roles through education, training, and collaborations that promote participation and opportunity.

IV. Roles & Responsibilities

Several roles are responsible for enforcing this policy, outlined below.

- The [Chief Information Officer (CIO), or equivalent] is responsible for directing [Agency] technology resources, policies, projects, services, and coordinating the same with other [Agency] departments. The [CIO, or equivalent] shall designate the [Chief Information Security Officer (CISO), or equivalent] to actively ensure AI systems are used in accordance with the [Agency Information and System Security Policy, or applicable policy]. The [CIO, or equivalent] shall also designate the [Chief AI Officer (CAIO), or equivalent].

- The [CAIO, or equivalent] is responsible for ensuring that the [Agency's] use of AI systems is done so in accordance with this policy and the [other related Agency policy].

- The [CISO, or equivalent] is responsible for overseeing the enterprise security infrastructure, cybersecurity operations, updating security policies, procedures, standards, guidelines, and monitoring policy compliance.

- The [Chief Privacy Officer (CPO), or equivalent] is responsible for overseeing the enterprise digital privacy practices, data processing practices, and responsible usage of technology in compliance with the [Agency Privacy Policy, or applicable policy].

- [Agency] departments are responsible for following this policy and following updates to this policy and the [AI Governance Manual, or applicable policy], and shall check compliance with these documents at least annually.

- The [CIO or designee, or equivalent] shall notify [Agency] departments when an update to this policy or the [AI Governance Manual, or applicable policy] is released.

- The [Attorney's Office, or equivalent] is responsible for advising of any legal issues or risks associated with AI systems usage by or on behalf of [Agency] departments.

- The [Executive Office or designee, or equivalent] may, at its discretion, inspect the usage of AI systems and require a department to alter or cease its usage of AI systems or a partner's usage of AI systems on behalf of the department.

- The [Finance Department – Purchasing Office, or equivalent] is responsible for overseeing the procurement of AI systems in alignment with existing IT procurement and governance processes and requiring vendors to comply with [Agency] policy standards through contractual agreements.

V. Policy

When purchasing, configuring, developing, operating, or maintaining AI systems, the [Agency]

will:

1. Uphold the Guiding Principles for Responsible AI Systems;

2. Conduct an AI Review to assess the potential risk of AI systems. The [CAIO, or equivalent] is responsible for coordinating review of AI systems used by the [Agency] as detailed in the [AI Governance Manual, or applicable policy];

3. Obtain technical documentation about AI systems using the AI FactSheet or create equivalent documentation if internally developing the AI system. The [Finance Department, Purchasing Office, or equivalent] is responsible for requiring vendors to complete the AI FactSheet;

4. Require contractors to comply with the [Requirements for AI Systems, or applicable legal addendum] overseen by the [Finance Department, Purchasing Office or equivalent]; and

5. In the event of an incident involving the use of the AI system, the [Agency] will follow an Incident Response Plan as detailed in the [AI Incident Response Plan, or equivalent]. The [CISO, or equivalent] is responsible for overseeing the security practices of AI systems used by or on behalf of [Agency] departments.

Prohibited Uses

The use of certain AI systems is prohibited due to the sensitive nature of the information processed and severe potential risk. This includes the following prohibited purposes:

- [Real-time and covert biometric identification.]
- Classification of human facial or body movements into certain emotions or sentiment with the use of computer vision techniques or emotion analysis. (e.g., positive, negative, neutral, happy, angry, nervous).]
- [Fully automated decisions that do not require any meaningful human oversight but substantially impact individuals.]
- [Social scoring, or the use of AI systems to track and classify individuals based on their behaviors, socioeconomic status, or personal characteristics.]
- [Cognitive behavioral manipulation of people or specific vulnerable groups.]
- [Autonomous weapons systems.]

If [Agency] staff become aware of an instance where an AI system has caused harm, staff must report the instance to their supervisor and the [CAIO, or equivalent].

Sunset Procedures

If an AI system operated by the [Agency] or on its behalf ceases to provide a positive utility to

the [Agency's] residents as determined by the [CAIO, or equivalent], then the use of that AI system must be halted unless express exception is provided by the [City Manager or City Council, or equivalents]. If the abrupt cessation of the use of that AI system would significantly disrupt the delivery of [Agency] services, usage of the AI system shall be gradually phased out over time.

Public Records
The [Agency] is subject to the [applicable public records act policy]. [Agency] staff must follow all current procedures for records retention and disclosure.

Policy Enforcement
All employees and agents of the [Agency], whether permanent or temporary, interns, volunteers, contractors, consultants, vendors, and other third parties operating AI systems on behalf of the [Agency] are required to abide by this Policy and the associated [AI Governance Manual, or applicable policy].

VI. Violations of the AI Policy

Violations of any section of the AI Policy, including failure to comply with the [Agency's] [AI Governance Manual, or applicable handbook], may be subject to disciplinary action, up to and including termination. Violations made by a third party while operating an AI system on behalf of the [Agency] may result in a breach of contract and/or pursuit of damages. Infractions that violate local, state, federal or international law may be remanded to the proper authorities.

VII. Exceptions (optional)

Staff, as described in the scope of this policy, must adhere to all controls outlined in this policy document unless a specific documented exception is explicitly granted by [CAIO, or equivalent]. Policy violations that have not been formally documented as an exception will be treated as a security incident in accordance with [Agency] policy. [Agency] staff or Departments may encounter emergency situations that necessitate the immediate use of AI tools or technologies. In these circumstances, each incident use case must obtain documented approval via the [Department Director, or equivalent] prior to use. In such situations, the [Department, staff or equivalent] must submit the [Department] approved use case to the [CAIO, or equivalent] for review within [30 days, or other duration] of the incident.

Restricted Platform Alternatives (optional)
In the event that the [Agency] restricts access to certain AI platforms, whenever practicable, the [Agency] will provide a URL redirect to a [Agency intranet website] that will provide the employee with alternative AI platforms which have been vetted and approved by the [Agency] for use.

The purpose of this policy is to help prevent the exfiltration of [Agency] data to personal devices which would violate [Agency Acceptable Use Policies, Privacy Policy, Information Security Policy]. This approach allows employees to choose amongst various platforms which

meet applicable privacy and security standards established by the [Agency] to facilitate productivity.

The [CISO, or equivalent] will designate a department to perform regular assessments of AI platforms and to regularly monitor such platforms so that the [Agency] can restrict and remove access in the event of a security or privacy incident.

VIII. Terms & Definitions

Artificial Intelligence: "Artificial intelligence" or "AI" to be a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action.

Algorithm: A series of logical steps through which an agent (typically a computer or software program) turns particular inputs into particular outputs.

AI system: Any system, software, sensor, or process that automatically generates outputs including, but not limited to, predictions, recommendations, or decisions that augment or replace human decision-making. This extends to software, hardware, algorithms, and data generated by these systems, used to automate large-scale processes or analyze large data sets.

Additional terms and definitions are provided in the AI Governance Manual.

IX. Acknowledgement

This policy was developed through the coordinated efforts of thousands in the GovAI Coalition who empower governments to leverage AI for the public good.

Approved:

_____ _____

[Chief Information Officer or equivalent] Date

Approved for posting:

_____ _____

[Director of Employee Relations Director of Human Resources, or equivalents] Date

NIST AI 100-1

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST AI 100-1

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.100-1

January 2023

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

This publication is available free of charge from: https://doi.org/10.6028/NIST.AI.100-1

Update Schedule and Versions

The Artificial Intelligence Risk Management Framework (AI RMF) is intended to be a living document.

NIST will review the content and usefulness of the Framework regularly to determine if an update is appropriate; a review with formal input from the AI community is expected to take place no later than 2028. The Framework will employ a two-number versioning system to track and identify major and minor changes. The first number will represent the generation of the AI RMF and its companion documents (e.g., 1.0) and will change only with major revisions. Minor revisions will be tracked using ".n" after the generation number (e.g., 1.1). All changes will be tracked using a Version Control Table which identifies the history, including version number, date of change, and description of change. NIST plans to update the AI RMF Playbook frequently. Comments on the AI RMF Playbook may be sent via email to AIframework@nist.gov at any time and will be reviewed and integrated on a semi-annual basis.

Table of Contents

List of Tables

List of Figures

Executive Summary

Artificial intelligence (AI) technologies have significant potential to transform society and people's lives – from commerce and health to transportation and cybersecurity to the environment and our planet. AI technologies can drive inclusive economic growth and support scientific advancements that improve the conditions of our world. AI technologies, however, also pose risks that can negatively impact individuals, groups, organizations, communities, society, the environment, and the planet. Like risks for other types of technology, AI risks can emerge in a variety of ways and can be characterized as long- or short-term, high- or low-probability, systemic or localized, and high- or low-impact.

> The AI RMF refers to an *AI system* as an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy (Adapted from: OECD Recommendation on AI:2019; ISO/IEC 22989:2022).

While there are myriad standards and best practices to help organizations mitigate the risks of traditional software or information-based systems, the risks posed by AI systems are in many ways unique (See Appendix B). AI systems, for example, may be trained on data that can change over time, sometimes significantly and unexpectedly, affecting system functionality and trustworthiness in ways that are hard to understand. AI systems and the contexts in which they are deployed are frequently complex, making it difficult to detect and respond to failures when they occur. AI systems are inherently socio-technical in nature, meaning they are influenced by societal dynamics and human behavior. AI risks – and benefits – can emerge from the interplay of technical aspects combined with societal factors related to how a system is used, its interactions with other AI systems, who operates it, and the social context in which it is deployed.

These risks make AI a uniquely challenging technology to deploy and utilize both for organizations and within society. Without proper controls, AI systems can amplify, perpetuate, or exacerbate inequitable or undesirable outcomes for individuals and communities. With proper controls, AI systems can mitigate and manage inequitable outcomes.

AI risk management is a key component of responsible development and use of AI systems. Responsible AI practices can help align the decisions about AI system design, development, and uses with intended aim and values. Core concepts in responsible AI emphasize human centricity, social responsibility, and sustainability. AI risk management can drive responsible uses and practices by prompting organizations and their internal teams who design, develop, and deploy AI to think more critically about context and potential or unexpected negative and positive impacts. Understanding and managing the risks of AI systems will help to enhance trustworthiness, and in turn, cultivate public trust.

Social responsibility can refer to the organization's responsibility "for the impacts of its decisions and activities on society and the environment through transparent and ethical behavior" (ISO 26000:2010). *Sustainability* refers to the "state of the global system, including environmental, social, and economic aspects, in which the needs of the present are met without compromising the ability of future generations to meet their own needs" (ISO/IEC TR 24368:2022). Responsible AI is meant to result in technology that is also equitable and accountable. The expectation is that organizational practices are carried out in accord with "*professional responsibility*," defined by ISO as an approach that "aims to ensure that professionals who design, develop, or deploy AI systems and applications or AI-based products or systems, recognize their unique position to exert influence on people, society, and the future of AI" (ISO/IEC TR 24368:2022).

As directed by the National Artificial Intelligence Initiative Act of 2020 (P.L. 116-283), the goal of the AI RMF is to offer a resource to the organizations designing, developing, deploying, or using AI systems to help manage the many risks of AI and promote trustworthy and responsible development and use of AI systems. The Framework is intended to be ***voluntary***, rights-preserving, non-sector-specific, and use-case agnostic, providing flexibility to organizations of all sizes and in all sectors and throughout society to implement the approaches in the Framework.

The Framework is designed to equip organizations and individuals – referred to here as *AI actors* – with approaches that increase the trustworthiness of AI systems, and to help foster the responsible design, development, deployment, and use of AI systems over time. AI actors are defined by the Organisation for Economic Co-operation and Development (OECD) as "those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI" [OECD (2019) Artificial Intelligence in Society—OECD iLibrary] (See Appendix A).

The AI RMF is intended to be practical, to adapt to the AI landscape as AI technologies continue to develop, and to be operationalized by organizations in varying degrees and capacities so society can benefit from AI while also being protected from its potential harms.

The Framework and supporting resources will be updated, expanded, and improved based on evolving technology, the standards landscape around the world, and AI community experience and feedback. NIST will continue to align the AI RMF and related guidance with applicable international standards, guidelines, and practices. As the AI RMF is put into use, additional lessons will be learned to inform future updates and additional resources.

The Framework is divided into two parts. Part 1 discusses how organizations can frame the risks related to AI and describes the intended audience. Next, AI risks and trustworthiness are analyzed, outlining the characteristics of trustworthy AI systems, which include

valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy enhanced, and fair with their harmful biases managed.

Part 2 comprises the "Core" of the Framework. It describes four specific functions to help organizations address the risks of AI systems in practice. These functions – **GOVERN**, **MAP**, **MEASURE**, and **MANAGE** – are broken down further into categories and subcategories. While **GOVERN** applies to all stages of organizations' AI risk management processes and procedures, the **MAP**, **MEASURE**, and **MANAGE** functions can be applied in AI system-specific contexts and at specific stages of the AI lifecycle.

Additional resources related to the Framework are included in the AI RMF Playbook, which is available via the NIST AI RMF website:
https://www.nist.gov/itl/ai-risk-management-framework.

Development of the AI RMF by NIST in collaboration with the private and public sectors is directed and consistent with its broader AI efforts called for by the National AI Initiative Act of 2020, the National Security Commission on Artificial Intelligence recommendations, and the Plan for Federal Engagement in Developing Technical Standards and Related Tools. Engagement with the AI community during this Framework's development – via responses to a formal Request for Information, three widely attended workshops, public comments on a concept paper and two drafts of the Framework, discussions at multiple public forums, and many small group meetings – has informed development of the AI RMF 1.0 as well as AI research and development and evaluation conducted by NIST and others. Priority research and additional guidance that will enhance this Framework will be captured in an associated AI Risk Management Framework Roadmap to which NIST and the broader community can contribute.

Part 1: Foundational Information

1. Framing Risk

AI risk management offers a path to minimize potential negative impacts of AI systems, such as threats to civil liberties and rights, while also providing opportunities to maximize positive impacts. Addressing, documenting, and managing AI risks and potential negative impacts effectively can lead to more trustworthy AI systems.

1.1 Understanding and Addressing Risks, Impacts, and Harms

In the context of the AI RMF, *risk* refers to the composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or threats (Adapted from: ISO 31000:2018). When considering the negative impact of a potential event, risk is a function of 1) the negative impact, or magnitude of harm, that would arise if the circumstance or event occurs and 2) the likelihood of occurrence (Adapted from: OMB Circular A-130:2016). Negative impact or harm can be experienced by individuals, groups, communities, organizations, society, the environment, and the planet.

> "Risk management refers to coordinated activities to direct and control an organization with regard to risk" (Source: ISO 31000:2018).

While risk management processes generally address negative impacts, this Framework offers approaches to minimize anticipated negative impacts of AI systems *and* identify opportunities to maximize positive impacts. Effectively managing the risk of potential harms could lead to more trustworthy AI systems and unleash potential benefits to people (individuals, communities, and society), organizations, and systems/ecosystems. Risk management can enable AI developers and users to understand impacts and account for the inherent limitations and uncertainties in their models and systems, which in turn can improve overall system performance and trustworthiness and the likelihood that AI technologies will be used in ways that are beneficial.

The AI RMF is designed to address new risks as they emerge. This flexibility is particularly important where impacts are not easily foreseeable and applications are evolving. While some AI risks and benefits are well-known, it can be challenging to assess negative impacts and the degree of harms. Figure 1 provides examples of potential harms that can be related to AI systems.

AI risk management efforts should consider that humans may assume that AI systems work – and work well – in *all* settings. For example, whether correct or not, AI systems are often perceived as being more objective than humans or as offering greater capabilities than general software.

Fig. 1. Examples of potential harms related to AI systems. Trustworthy AI systems and their responsible use can mitigate negative risks and contribute to benefits for people, organizations, and ecosystems.

1.2 Challenges for AI Risk Management

Several challenges are described below. They should be taken into account when managing risks in pursuit of AI trustworthiness.

1.2.1 Risk Measurement

AI risks or failures that are not well-defined or adequately understood are difficult to measure quantitatively or qualitatively. The inability to appropriately measure AI risks does not imply that an AI system necessarily poses either a high or low risk. Some risk measurement challenges include:

Risks related to third-party software, hardware, and data: Third-party data or systems can accelerate research and development and facilitate technology transition. They also may complicate risk measurement. Risk can emerge both from third-party data, software or hardware itself and how it is used. Risk metrics or methodologies used by the organization developing the AI system may not align with the risk metrics or methodologies uses by the organization *deploying or operating* the system. Also, the organization developing the AI system may not be transparent about the risk metrics or methodologies it used. Risk measurement and management can be complicated by how customers use or integrate third-party data or systems into AI products or services, particularly without sufficient internal governance structures and technical safeguards. Regardless, all parties and AI actors should manage risk in the AI systems they develop, deploy, or use as standalone or integrated components.

Tracking emergent risks: Organizations' risk management efforts will be enhanced by identifying and tracking emergent risks and considering techniques for measuring them.

AI system impact assessment approaches can help AI actors understand potential impacts or harms within specific contexts.

Availability of reliable metrics: The current lack of consensus on robust and verifiable measurement methods for risk and trustworthiness, and applicability to different AI use cases, is an AI risk measurement challenge. Potential pitfalls when seeking to measure negative risk or harms include the reality that development of metrics is often an institutional endeavor and may inadvertently reflect factors unrelated to the underlying impact. In addition, measurement approaches can be oversimplified, gamed, lack critical nuance, become relied upon in unexpected ways, or fail to account for differences in affected groups and contexts.

Approaches for measuring impacts on a population work best if they recognize that contexts matter, that harms may affect varied groups or sub-groups differently, and that communities or other sub-groups who may be harmed are not always direct users of a system.

Risk at different stages of the AI lifecycle: Measuring risk at an earlier stage in the AI lifecycle may yield different results than measuring risk at a later stage; some risks may be latent at a given point in time and may increase as AI systems adapt and evolve. Furthermore, different AI actors across the AI lifecycle can have different risk perspectives. For example, an AI developer who makes AI software available, such as pre-trained models, can have a different risk perspective than an AI actor who is responsible for deploying that pre-trained model in a specific use case. Such deployers may not recognize that their particular uses could entail risks which differ from those perceived by the initial developer. All involved AI actors share responsibilities for designing, developing, and deploying a trustworthy AI system that is fit for purpose.

Risk in real-world settings: While measuring AI risks in a laboratory or a controlled environment may yield important insights pre-deployment, these measurements may differ from risks that emerge in operational, real-world settings.

Inscrutability: Inscrutable AI systems can complicate risk measurement. Inscrutability can be a result of the opaque nature of AI systems (limited explainability or interpretability), lack of transparency or documentation in AI system development or deployment, or inherent uncertainties in AI systems.

Human baseline: Risk management of AI systems that are intended to augment or replace human activity, for example decision making, requires some form of baseline metrics for comparison. This is difficult to systematize since AI systems carry out different tasks – and perform tasks differently – than humans.

1.2.2 Risk Tolerance

While the AI RMF can be used to prioritize risk, it does not prescribe risk tolerance. *Risk tolerance* refers to the organization's or AI actor's (see Appendix A) readiness to bear the risk in order to achieve its objectives. Risk tolerance can be influenced by legal or regulatory requirements (Adapted from: ISO GUIDE 73). Risk tolerance and the level of risk that is acceptable to organizations or society are highly contextual and application and use-case specific. Risk tolerances can be influenced by policies and norms established by AI system owners, organizations, industries, communities, or policy makers. Risk tolerances are likely to change over time as AI systems, policies, and norms evolve. Different organizations may have varied risk tolerances due to their particular organizational priorities and resource considerations.

Emerging knowledge and methods to better inform harm/cost-benefit tradeoffs will continue to be developed and debated by businesses, governments, academia, and civil society. To the extent that challenges for specifying AI risk tolerances remain unresolved, there may be contexts where a risk management framework is not yet readily applicable for mitigating negative AI risks.

The Framework is intended to be flexible and to augment existing risk practices which should align with applicable laws, regulations, and norms. Organizations should follow existing regulations and guidelines for risk criteria, tolerance, and response established by organizational, domain, discipline, sector, or professional requirements. Some sectors or industries may have established definitions of harm or established documentation, reporting, and disclosure requirements. Within sectors, risk management may depend on existing guidelines for specific applications and use case settings. Where established guidelines do not exist, organizations should define reasonable risk tolerance. Once tolerance is defined, this AI RMF can be used to manage risks and to document risk management processes.

1.2.3 Risk Prioritization

Attempting to eliminate negative risk entirely can be counterproductive in practice because not all incidents and failures can be eliminated. Unrealistic expectations about risk may lead organizations to allocate resources in a manner that makes risk triage inefficient or impractical or wastes scarce resources. A risk management culture can help organizations recognize that not all AI risks are the same, and resources can be allocated purposefully. Actionable risk management efforts lay out clear guidelines for assessing trustworthiness of each AI system an organization develops or deploys. Policies and resources should be prioritized based on the assessed risk level and potential impact of an AI system. The extent to which an AI system may be customized or tailored to the specific context of use by the AI deployer can be a contributing factor.

When applying the AI RMF, risks which the organization determines to be highest for the AI systems within a given context of use call for the most urgent prioritization and most thorough risk management process. In cases where an AI system presents unacceptable negative risk levels – such as where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present – development and deployment should cease in a safe manner until risks can be sufficiently managed. If an AI system's development, deployment, and use cases are found to be low-risk in a specific context, that may suggest potentially lower prioritization.

Risk prioritization may differ between AI systems that are designed or deployed to directly interact with humans as compared to AI systems that are not. Higher initial prioritization may be called for in settings where the AI system is trained on large datasets comprised of sensitive or protected data such as personally identifiable information, or where the outputs of the AI systems have direct or indirect impact on humans. AI systems designed to interact only with computational systems and trained on non-sensitive datasets (for example, data collected from the physical environment) may call for lower initial prioritization. Nonetheless, regularly assessing and prioritizing risk based on context remains important because non-human-facing AI systems can have downstream safety or social implications.

Residual risk – defined as risk remaining after risk treatment (Source: ISO GUIDE 73) – directly impacts end users or affected individuals and communities. Documenting residual risks will call for the system provider to fully consider the risks of deploying the AI product and will inform end users about potential negative impacts of interacting with the system.

1.2.4 Organizational Integration and Management of Risk

AI risks should not be considered in isolation. Different AI actors have different responsibilities and awareness depending on their roles in the lifecycle. For example, organizations developing an AI system often will not have information about how the system may be used. AI risk management should be integrated and incorporated into broader enterprise risk management strategies and processes. Treating AI risks along with other critical risks, such as cybersecurity and privacy, will yield a more integrated outcome and organizational efficiencies.

The AI RMF may be utilized along with related guidance and frameworks for managing AI system risks or broader enterprise risks. Some risks related to AI systems are common across other types of software development and deployment. Examples of overlapping risks include: privacy concerns related to the use of underlying data to train AI systems; the energy and environmental implications associated with resource-heavy computing demands; security concerns related to the confidentiality, integrity, and availability of the system and its training and output data; and general security of the underlying software and hardware for AI systems.

Organizations need to establish and maintain the appropriate accountability mechanisms, roles and responsibilities, culture, and incentive structures for risk management to be effective. Use of the AI RMF alone will not lead to these changes or provide the appropriate incentives. Effective risk management is realized through organizational commitment at senior levels and may require cultural change within an organization or industry. In addition, small to medium-sized organizations managing AI risks or implementing the AI RMF may face different challenges than large organizations, depending on their capabilities and resources.

2. Audience

Identifying and managing AI risks and potential impacts – both positive and negative – requires a broad set of perspectives and actors across the AI lifecycle. Ideally, AI actors will represent a diversity of experience, expertise, and backgrounds and comprise demographically and disciplinarily diverse teams. The AI RMF is intended to be used by AI actors across the AI lifecycle and dimensions.

The OECD has developed a framework for classifying AI lifecycle activities according to five key socio-technical dimensions, each with properties relevant for AI policy and governance, including risk management [OECD (2022) OECD Framework for the Classification of AI systems — OECD Digital Economy Papers]. Figure 2 shows these dimensions, slightly modified by NIST for purposes of this framework. The NIST modification highlights the importance of test, evaluation, verification, and validation (TEVV) processes throughout an AI lifecycle and generalizes the operational context of an AI system.

AI dimensions displayed in Figure 2 are the Application Context, Data and Input, AI Model, and Task and Output. AI actors involved in these dimensions who perform or manage the design, development, deployment, evaluation, and use of AI systems and drive AI risk management efforts are the *primary* AI RMF audience.

Representative AI actors across the lifecycle dimensions are listed in Figure 3 and described in detail in Appendix A. Within the AI RMF, all AI actors work together to manage risks and achieve the goals of trustworthy and responsible AI. AI actors with TEVV-specific expertise are integrated throughout the AI lifecycle and are especially likely to benefit from the Framework. Performed regularly, TEVV tasks can provide insights relative to technical, societal, legal, and ethical standards or norms, and can assist with anticipating impacts and assessing and tracking emergent risks. As a regular process within an AI lifecycle, TEVV allows for both mid-course remediation and post-hoc risk management.

The People & Planet dimension at the center of Figure 2 represents human rights and the broader well-being of society and the planet. The AI actors in this dimension comprise a separate AI RMF audience who *informs* the primary audience. These AI actors may include trade associations, standards developing organizations, researchers, advocacy groups,

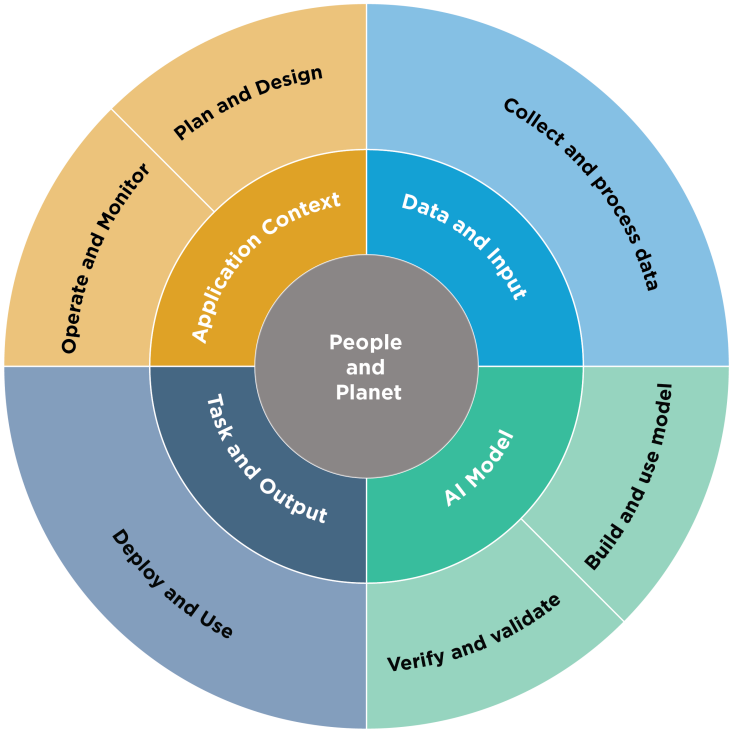

Fig. 2. Lifecycle and Key Dimensions of an AI System. Modified from OECD (2022) OECD Framework for the Classification of AI systems — OECD Digital Economy Papers. The two inner circles show AI systems' key dimensions and the outer circle shows AI lifecycle stages. Ideally, risk management efforts start with the Plan and Design function in the application context and are performed throughout the AI system lifecycle. See Figure 3 for representative AI actors.

environmental groups, civil society organizations, end users, and potentially impacted individuals and communities. These actors can:

- assist in providing context and understanding potential and actual impacts;
- be a source of formal or quasi-formal norms and guidance for AI risk management;
- designate boundaries for AI operation (technical, societal, legal, and ethical); and
- promote discussion of the tradeoffs needed to balance societal values and priorities related to civil liberties and rights, equity, the environment and the planet, and the economy.

Successful risk management depends upon a sense of collective responsibility among AI actors shown in Figure 3. The AI RMF functions, described in Section 5, require diverse perspectives, disciplines, professions, and experiences. Diverse teams contribute to more open sharing of ideas and assumptions about the purposes and functions of technology – making these implicit aspects more explicit. This broader collective perspective creates opportunities for surfacing problems and identifying existing and emergent risks.

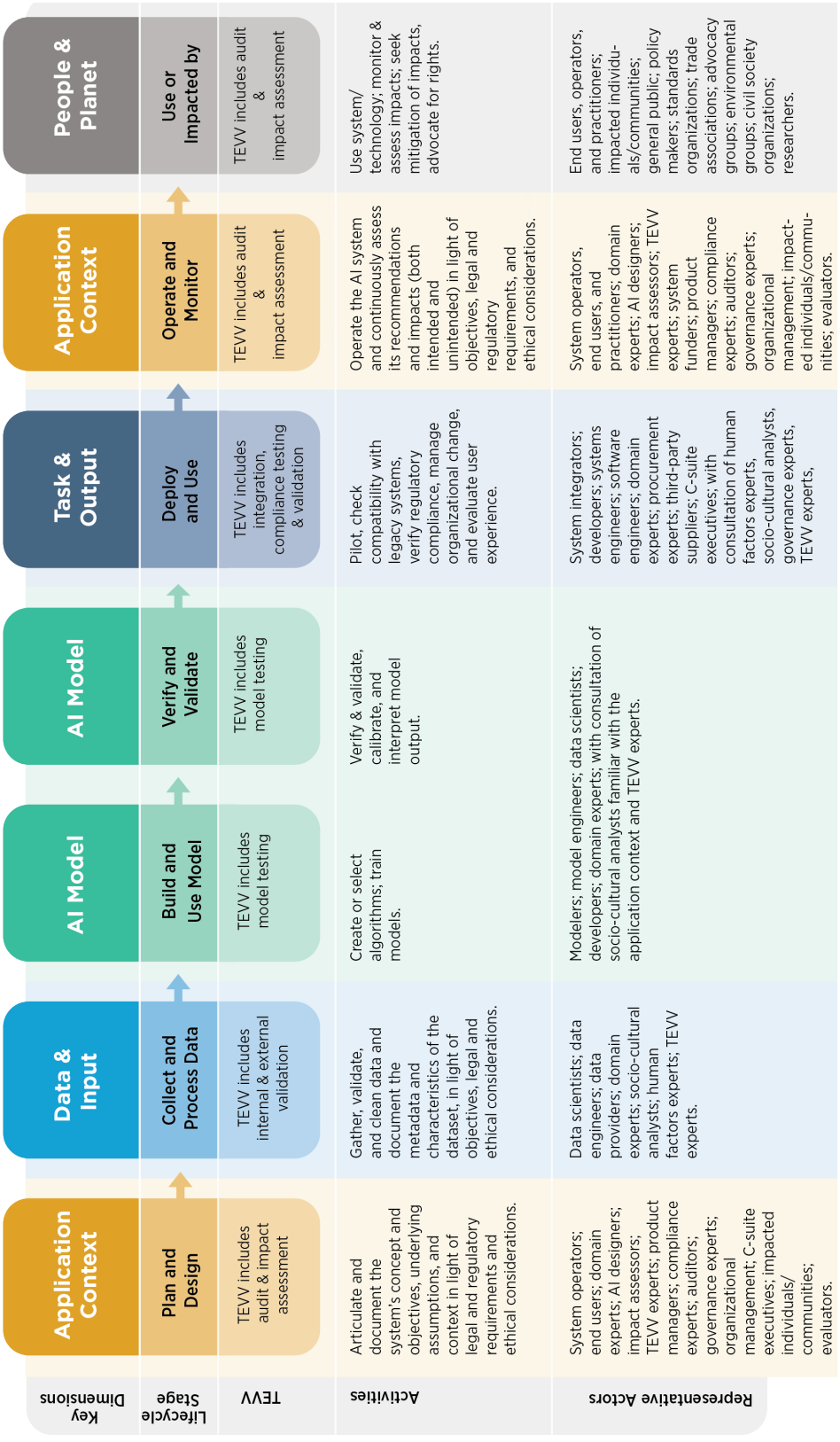

Fig. 3. AI actors across AI lifecycle stages. See Appendix A for detailed descriptions of AI actor tasks, including details about testing, evaluation, verification, and validation tasks. Note that AI actors in the AI Model dimension (Figure 2) are separated as a best practice, with those building and using the models separated from those verifying and validating the models.

3. AI Risks and Trustworthiness

For AI systems to be trustworthy, they often need to be responsive to a multiplicity of criteria that are of value to interested parties. Approaches which enhance AI trustworthiness can reduce negative AI risks. This Framework articulates the following **characteristics** of trustworthy AI and offers guidance for addressing them. Characteristics of trustworthy AI systems include: **valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed.** Creating trustworthy AI requires balancing each of these characteristics based on the AI system's context of use. While all characteristics are socio-technical system attributes, accountability and transparency also relate to the processes and activities internal to an AI system and its external setting. Neglecting these characteristics can increase the probability and magnitude of negative consequences.

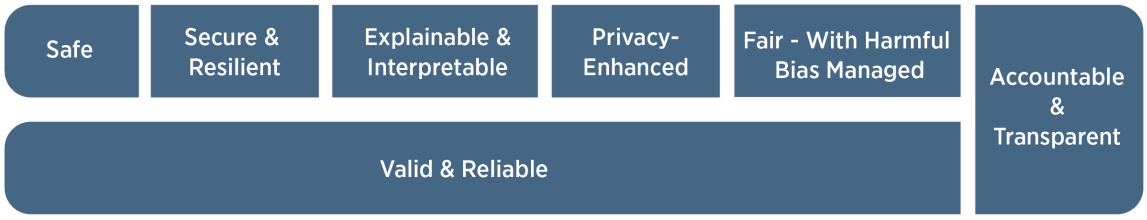

Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.

Trustworthiness characteristics (shown in Figure 4) are inextricably tied to social and organizational behavior, the datasets used by AI systems, selection of AI models and algorithms and the decisions made by those who build them, and the interactions with the humans who provide insight from and oversight of such systems. Human judgment should be employed when deciding on the specific metrics related to AI trustworthiness characteristics and the precise threshold values for those metrics.

Addressing AI trustworthiness characteristics individually will not ensure AI system trustworthiness; tradeoffs are usually involved, rarely do all characteristics apply in every setting, and some will be more or less important in any given situation. Ultimately, trustworthiness is a social concept that ranges across a spectrum and is only as strong as its weakest characteristics.

When managing AI risks, organizations can face difficult decisions in balancing these characteristics. For example, in certain scenarios tradeoffs may emerge between optimizing for interpretability and achieving privacy. In other cases, organizations might face a tradeoff between predictive accuracy and interpretability. Or, under certain conditions such as data sparsity, privacy-enhancing techniques can result in a loss in accuracy, affecting decisions

about fairness and other values in certain domains. Dealing with tradeoffs requires taking into account the decision-making context. These analyses can highlight the existence and extent of tradeoffs between different measures, but they do not answer questions about how to navigate the tradeoff. Those depend on the values at play in the relevant *context* and should be resolved in a manner that is both transparent and appropriately justifiable.

There are multiple approaches for enhancing contextual awareness in the AI lifecycle. For example, subject matter experts can assist in the evaluation of TEVV findings and work with product and deployment teams to align TEVV parameters to requirements and deployment conditions. When properly resourced, increasing the breadth and diversity of input from interested parties and relevant AI actors throughout the AI lifecycle can enhance opportunities for informing contextually sensitive evaluations, and for identifying AI system benefits and positive impacts. These practices can increase the likelihood that risks arising in social contexts are managed appropriately.

Understanding and treatment of trustworthiness characteristics depends on an AI actor's particular role within the AI lifecycle. For any given AI system, an AI designer or developer may have a different perception of the characteristics than the deployer.

Trustworthiness characteristics explained in this document influence each other. Highly secure but unfair systems, accurate but opaque and uninterpretable systems, and inaccurate but secure, privacy-enhanced, and transparent systems are all undesirable. A comprehensive approach to risk management calls for balancing tradeoffs among the trustworthiness characteristics. It is the joint responsibility of all AI actors to determine whether AI technology is an appropriate or necessary tool for a given context or purpose, and how to use it responsibly. The decision to commission or deploy an AI system should be based on a contextual assessment of trustworthiness characteristics and the relative risks, impacts, costs, and benefits, and informed by a broad set of interested parties.

3.1 Valid and Reliable

Validation is the "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled" (Source: ISO 9000:2015). Deployment of AI systems which are inaccurate, unreliable, or poorly generalized to data and settings beyond their training creates and increases negative AI risks and reduces trustworthiness.

Reliability is defined in the same standard as the "ability of an item to perform as required, without failure, for a given time interval, under given conditions" (Source: ISO/IEC TS 5723:2022). Reliability is a goal for overall correctness of AI system operation under the conditions of expected use and over a given period of time, including the entire lifetime of the system.

Accuracy and robustness contribute to the validity and trustworthiness of AI systems, and can be in tension with one another in AI systems.

Accuracy is defined by ISO/IEC TS 5723:2022 as "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true." Measures of accuracy should consider computational-centric measures (e.g., false positive and false negative rates), human-AI teaming, and demonstrate external validity (generalizable beyond the training conditions). Accuracy measurements should always be paired with clearly defined and realistic test sets – that are representative of conditions of expected use – and details about test methodology; these should be included in associated documentation. Accuracy measurements may include disaggregation of results for different data segments.

Robustness or *generalizability* is defined as the "ability of a system to maintain its level of performance under a variety of circumstances" (Source: ISO/IEC TS 5723:2022). Robustness is a goal for appropriate system functionality in a broad set of conditions and circumstances, including uses of AI systems not initially anticipated. Robustness requires not only that the system perform exactly as it does under expected uses, but also that it should perform in ways that minimize potential harms to people if it is operating in an unexpected setting.

Validity and reliability for deployed AI systems are often assessed by ongoing testing or monitoring that confirms a system is performing as intended. Measurement of validity, accuracy, robustness, and reliability contribute to trustworthiness and should take into consideration that certain types of failures can cause greater harm. AI risk management efforts should prioritize the minimization of potential negative impacts, and may need to include human intervention in cases where the AI system cannot detect or correct errors.

3.2 Safe

AI systems should "not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered" (Source: ISO/IEC TS 5723:2022). Safe operation of AI systems is improved through:

- responsible design, development, and deployment practices;
- clear information to deployers on responsible use of the system;
- responsible decision-making by deployers and end users; and
- explanations and documentation of risks based on empirical evidence of incidents.

Different types of safety risks may require tailored AI risk management approaches based on context and the severity of potential risks presented. Safety risks that pose a potential risk of serious injury or death call for the most urgent prioritization and most thorough risk management process.

Employing safety considerations during the lifecycle and starting as early as possible with planning and design can prevent failures or conditions that can render a system dangerous. Other practical approaches for AI safety often relate to rigorous simulation and in-domain testing, real-time monitoring, and the ability to shut down, modify, or have human intervention into systems that deviate from intended or expected functionality.

AI safety risk management approaches should take cues from efforts and guidelines for safety in fields such as transportation and healthcare, and align with existing sector- or application-specific guidelines or standards.

3.3 Secure and Resilient

AI systems, as well as the ecosystems in which they are deployed, may be said to be *resilient* if they can withstand unexpected adverse events or unexpected changes in their environment or use – or if they can maintain their functions and structure in the face of internal and external change and degrade safely and gracefully when this is necessary (Adapted from: ISO/IEC TS 5723:2022). Common security concerns relate to adversarial examples, data poisoning, and the exfiltration of models, training data, or other intellectual property through AI system endpoints. AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use may be said to be *secure*. Guidelines in the NIST Cybersecurity Framework and Risk Management Framework are among those which are applicable here.

Security and resilience are related but distinct characteristics. While resilience is the ability to return to normal function after an unexpected adverse event, security includes resilience but also encompasses protocols to avoid, protect against, respond to, or recover from attacks. Resilience relates to robustness and goes beyond the provenance of the data to encompass unexpected or adversarial use (or abuse or misuse) of the model or data.

3.4 Accountable and Transparent

Trustworthy AI depends upon accountability. Accountability presupposes transparency. *Transparency* reflects the extent to which information about an AI system and its outputs is available to individuals interacting with such a system – regardless of whether they are even aware that they are doing so. Meaningful transparency provides access to appropriate levels of information based on the stage of the AI lifecycle and tailored to the role or knowledge of AI actors or individuals interacting with or using the AI system. By promoting higher levels of understanding, transparency increases confidence in the AI system.

This characteristic's scope spans from design decisions and training data to model training, the structure of the model, its intended use cases, and how and when deployment, post-deployment, or end user decisions were made and by whom. Transparency is often necessary for actionable redress related to AI system outputs that are incorrect or otherwise lead to negative impacts. Transparency should consider human-AI interaction: for exam-

ple, how a human operator or user is notified when a potential or actual adverse outcome caused by an AI system is detected. A transparent system is not necessarily an accurate, privacy-enhanced, secure, or fair system. However, it is difficult to determine whether an opaque system possesses such characteristics, and to do so over time as complex systems evolve.

The role of AI actors should be considered when seeking accountability for the outcomes of AI systems. The relationship between risk and accountability associated with AI and technological systems more broadly differs across cultural, legal, sectoral, and societal contexts. When consequences are severe, such as when life and liberty are at stake, AI developers and deployers should consider proportionally and proactively adjusting their transparency and accountability practices. Maintaining organizational practices and governing structures for harm reduction, like risk management, can help lead to more accountable systems.

Measures to enhance transparency and accountability should also consider the impact of these efforts on the implementing entity, including the level of necessary resources and the need to safeguard proprietary information.

Maintaining the provenance of training data and supporting attribution of the AI system's decisions to subsets of training data can assist with both transparency and accountability. Training data may also be subject to copyright and should follow applicable intellectual property rights laws.

As transparency tools for AI systems and related documentation continue to evolve, developers of AI systems are encouraged to test different types of transparency tools in cooperation with AI deployers to ensure that AI systems are used as intended.

3.5 Explainable and Interpretable

Explainability refers to a representation of the mechanisms underlying AI systems' operation, whereas *interpretability* refers to the meaning of AI systems' output in the context of their designed functional purposes. Together, explainability and interpretability assist those operating or overseeing an AI system, as well as users of an AI system, to gain deeper insights into the functionality and trustworthiness of the system, including its outputs. The underlying assumption is that perceptions of negative risk stem from a lack of ability to make sense of, or contextualize, system output appropriately. Explainable and interpretable AI systems offer information that will help end users understand the purposes and potential impact of an AI system.

Risk from lack of explainability may be managed by describing how AI systems function, with descriptions tailored to individual differences such as the user's role, knowledge, and skill level. Explainable systems can be debugged and monitored more easily, and they lend themselves to more thorough documentation, audit, and governance.

Risks to interpretability often can be addressed by communicating a description of why an AI system made a particular prediction or recommendation. (See "Four Principles of Explainable Artificial Intelligence" and "Psychological Foundations of Explainability and Interpretability in Artificial Intelligence" found here.)

Transparency, explainability, and interpretability are distinct characteristics that support each other. Transparency can answer the question of "what happened" in the system. Explainability can answer the question of "how" a decision was made in the system. Interpretability can answer the question of "why" a decision was made by the system and its meaning or context to the user.

3.6 Privacy-Enhanced

Privacy refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity. These norms and practices typically address freedom from intrusion, limiting observation, or individuals' agency to consent to disclosure or control of facets of their identities (e.g., body, data, reputation). (See The NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management.)

Privacy values such as anonymity, confidentiality, and control generally should guide choices for AI system design, development, and deployment. Privacy-related risks may influence security, bias, and transparency and come with tradeoffs with these other characteristics. Like safety and security, specific technical features of an AI system may promote or reduce privacy. AI systems can also present new risks to privacy by allowing inference to identify individuals or previously private information about individuals.

Privacy-enhancing technologies ("PETs") for AI, as well as data minimizing methods such as de-identification and aggregation for certain model outputs, can support design for privacy-enhanced AI systems. Under certain conditions such as data sparsity, privacy-enhancing techniques can result in a loss in accuracy, affecting decisions about fairness and other values in certain domains.

3.7 Fair – with Harmful Bias Managed

Fairness in AI includes concerns for equality and equity by addressing issues such as harmful bias and discrimination. Standards of fairness can be complex and difficult to define because perceptions of fairness differ among cultures and may shift depending on application. Organizations' risk management efforts will be enhanced by recognizing and considering these differences. Systems in which harmful biases are mitigated are not necessarily fair. For example, systems in which predictions are somewhat balanced across demographic groups may still be inaccessible to individuals with disabilities or affected by the digital divide or may exacerbate existing disparities or systemic biases.

Bias is broader than demographic balance and data representativeness. NIST has identified three major categories of AI bias to be considered and managed: systemic, computational and statistical, and human-cognitive. Each of these can occur in the absence of prejudice, partiality, or discriminatory intent. Systemic bias can be present in AI datasets, the organizational norms, practices, and processes across the AI lifecycle, and the broader society that uses AI systems. Computational and statistical biases can be present in AI datasets and algorithmic processes, and often stem from systematic errors due to non-representative samples. Human-cognitive biases relate to how an individual or group perceives AI system information to make a decision or fill in missing information, or how humans think about purposes and functions of an AI system. Human-cognitive biases are omnipresent in decision-making processes across the AI lifecycle and system use, including the design, implementation, operation, and maintenance of AI.

Bias exists in many forms and can become ingrained in the automated systems that help make decisions about our lives. While bias is not always a negative phenomenon, AI systems can potentially increase the speed and scale of biases and perpetuate and amplify harms to individuals, groups, communities, organizations, and society. Bias is tightly associated with the concepts of transparency as well as fairness in society. (For more information about bias, including the three categories, see NIST Special Publication 1270, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.)

4. Effectiveness of the AI RMF

Evaluations of AI RMF effectiveness – including ways to measure bottom-line improvements in the trustworthiness of AI systems – will be part of future NIST activities, in conjunction with the AI community.

Organizations and other users of the Framework are encouraged to periodically evaluate whether the AI RMF has improved their ability to manage AI risks, including but not limited to their policies, processes, practices, implementation plans, indicators, measurements, and expected outcomes. NIST intends to work collaboratively with others to develop metrics, methodologies, and goals for evaluating the AI RMF's effectiveness, and to broadly share results and supporting information. Framework users are expected to benefit from:

- enhanced processes for governing, mapping, measuring, and managing AI risk, and clearly documenting outcomes;
- improved awareness of the relationships and tradeoffs among trustworthiness characteristics, socio-technical approaches, and AI risks;
- explicit processes for making go/no-go system commissioning and deployment decisions;
- established policies, processes, practices, and procedures for improving organizational accountability efforts related to AI system risks;
- enhanced organizational culture which prioritizes the identification and management of AI system risks and potential impacts to individuals, communities, organizations, and society;
- better information sharing within and across organizations about risks, decision-making processes, responsibilities, common pitfalls, TEVV practices, and approaches for continuous improvement;
- greater contextual knowledge for increased awareness of downstream risks;
- strengthened engagement with interested parties and relevant AI actors; and
- augmented capacity for TEVV of AI systems and associated risks.

Part 2: Core and Profiles

5. AI RMF Core

The AI RMF Core provides outcomes and actions that enable dialogue, understanding, and activities to manage AI risks and responsibly develop trustworthy AI systems. As illustrated in Figure 5, the Core is composed of four functions: **GOVERN, MAP, MEASURE,** and **MANAGE**. Each of these high-level functions is broken down into categories and subcategories. Categories and subcategories are subdivided into specific actions and outcomes. Actions do not constitute a checklist, nor are they necessarily an ordered set of steps.

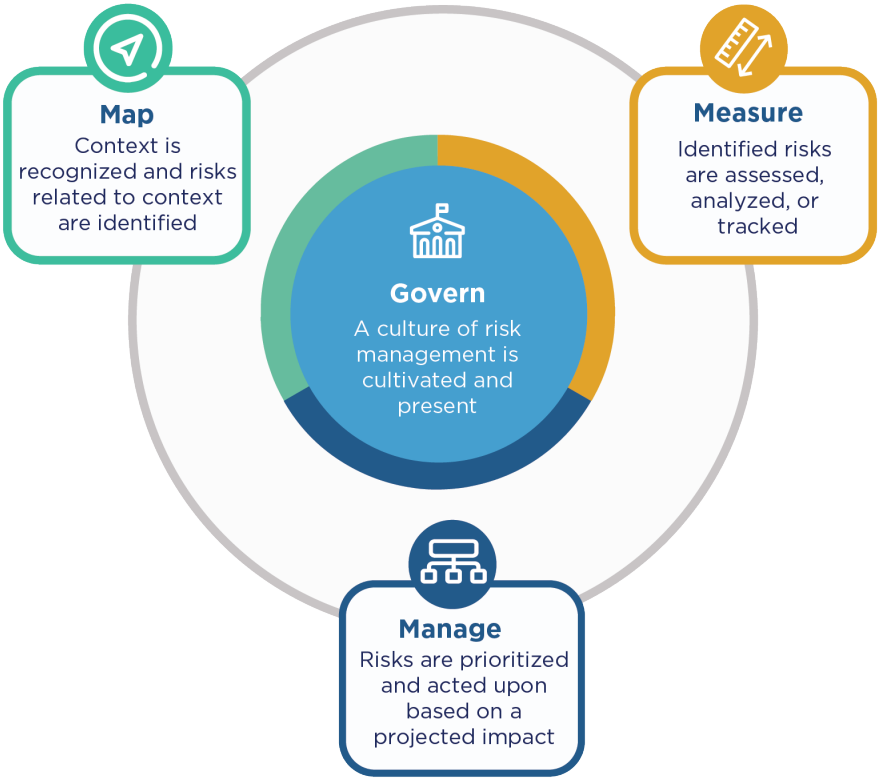

Fig. 5. Functions organize AI risk management activities at their highest level to govern, map, measure, and manage AI risks. Governance is designed to be a cross-cutting function to inform and be infused throughout the other three functions.

Risk management should be continuous, timely, and performed throughout the AI system lifecycle dimensions. AI RMF Core functions should be carried out in a way that reflects diverse and multidisciplinary perspectives, potentially including the views of AI actors outside the organization. Having a diverse team contributes to more open sharing of ideas and assumptions about purposes and functions of the technology being designed, developed,

deployed, or evaluated – which can create opportunities to surface problems and identify existing and emergent risks.

An online companion resource to the AI RMF, the NIST AI RMF Playbook, is available to help organizations navigate the AI RMF and achieve its outcomes through suggested tactical actions they can apply within their own contexts. Like the AI RMF, the Playbook is voluntary and organizations can utilize the suggestions according to their needs and interests. Playbook users can create tailored guidance selected from suggested material for their own use and contribute their suggestions for sharing with the broader community. Along with the AI RMF, the Playbook is part of the NIST Trustworthy and Responsible AI Resource Center.

Framework users may apply these functions as best suits their needs for managing AI risks based on their resources and capabilities. Some organizations may choose to select from among the categories and subcategories; others may choose and have the capacity to apply all categories and subcategories. Assuming a governance structure is in place, functions may be performed in any order across the AI lifecycle as deemed to add value by a user of the framework. After instituting the outcomes in **GOVERN**, most users of the AI RMF would start with the **MAP** function and continue to **MEASURE** or **MANAGE**. However users integrate the functions, the process should be iterative, with cross-referencing between functions as necessary. Similarly, there are categories and subcategories with elements that apply to multiple functions, or that logically should take place before certain subcategory decisions.

5.1 Govern

The **GOVERN** function:

- cultivates and implements a culture of risk management within organizations designing, developing, deploying, evaluating, or acquiring AI systems;
- outlines processes, documents, and organizational schemes that anticipate, identify, and manage the risks a system can pose, including to users and others across society – and procedures to achieve those outcomes;
- incorporates processes to assess potential impacts;
- provides a structure by which AI risk management functions can align with organizational principles, policies, and strategic priorities;
- connects technical aspects of AI system design and development to organizational values and principles, and enables organizational practices and competencies for the individuals involved in acquiring, training, deploying, and monitoring such systems; and
- addresses full product lifecycle and associated processes, including legal and other issues concerning use of third-party software or hardware systems and data.

GOVERN is a cross-cutting function that is infused throughout AI risk management and enables the other functions of the process. Aspects of **GOVERN**, especially those related to compliance or evaluation, should be integrated into each of the other functions. Attention to governance is a continual and intrinsic requirement for effective AI risk management over an AI system's lifespan and the organization's hierarchy.

Strong governance can drive and enhance internal practices and norms to facilitate organizational risk culture. Governing authorities can determine the overarching policies that direct an organization's mission, goals, values, culture, and risk tolerance. Senior leadership sets the tone for risk management within an organization, and with it, organizational culture. Management aligns the technical aspects of AI risk management to policies and operations. Documentation can enhance transparency, improve human review processes, and bolster accountability in AI system teams.

After putting in place the structures, systems, processes, and teams described in the **GOVERN** function, organizations should benefit from a purpose-driven culture focused on risk understanding and management. It is incumbent on Framework users to continue to execute the **GOVERN** function as knowledge, cultures, and needs or expectations from AI actors evolve over time.

Practices related to governing AI risks are described in the NIST AI RMF Playbook. Table 1 lists the **GOVERN** function's categories and subcategories.

Table 1: Categories and subcategories for the **GOVERN** function.

Categories	Subcategories
GOVERN 1: Policies, processes, procedures, and practices across the organization related to the mapping, measuring, and managing of AI risks are in place, transparent, and implemented effectively.	**GOVERN 1.1:** Legal and regulatory requirements involving AI are understood, managed, and documented. **GOVERN 1.2:** The characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices. **GOVERN 1.3:** Processes, procedures, and practices are in place to determine the needed level of risk management activities based on the organization's risk tolerance. **GOVERN 1.4:** The risk management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities.

Continued on next page

Table 1: Categories and subcategories for the **GOVERN** function. (Continued)

Categories	Subcategories
	GOVERN 1.5: Ongoing monitoring and periodic review of the risk management process and its outcomes are planned and organizational roles and responsibilities clearly defined, including determining the frequency of periodic review. **GOVERN 1.6:** Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities. **GOVERN 1.7:** Processes and procedures are in place for decommissioning and phasing out AI systems safely and in a manner that does not increase risks or decrease the organization's trustworthiness.
GOVERN 2: Accountability structures are in place so that the appropriate teams and individuals are empowered, responsible, and trained for mapping, measuring, and managing AI risks.	**GOVERN 2.1:** Roles and responsibilities and lines of communication related to mapping, measuring, and managing AI risks are documented and are clear to individuals and teams throughout the organization. **GOVERN 2.2:** The organization's personnel and partners receive AI risk management training to enable them to perform their duties and responsibilities consistent with related policies, procedures, and agreements. **GOVERN 2.3:** Executive leadership of the organization takes responsibility for decisions about risks associated with AI system development and deployment.
GOVERN 3: Workforce diversity, equity, inclusion, and accessibility processes are prioritized in the mapping, measuring, and managing of AI risks throughout the lifecycle.	**GOVERN 3.1:** Decision-making related to mapping, measuring, and managing AI risks throughout the lifecycle is informed by a diverse team (e.g., diversity of demographics, disciplines, experience, expertise, and backgrounds). **GOVERN 3.2:** Policies and procedures are in place to define and differentiate roles and responsibilities for human-AI configurations and oversight of AI systems.
GOVERN 4: Organizational teams are committed to a culture	**GOVERN 4.1:** Organizational policies and practices are in place to foster a critical thinking and safety-first mindset in the design, development, deployment, and uses of AI systems to minimize potential negative impacts.

Continued on next page

Table 1: Categories and subcategories for the **GOVERN** function. (Continued)

Categories	Subcategories
that considers and communicates AI risk.	**GOVERN 4.2:** Organizational teams document the risks and potential impacts of the AI technology they design, develop, deploy, evaluate, and use, and they communicate about the impacts more broadly. **GOVERN 4.3:** Organizational practices are in place to enable AI testing, identification of incidents, and information sharing.
GOVERN 5: Processes are in place for robust engagement with relevant AI actors.	**GOVERN 5.1:** Organizational policies and practices are in place to collect, consider, prioritize, and integrate feedback from those external to the team that developed or deployed the AI system regarding the potential individual and societal impacts related to AI risks. **GOVERN 5.2:** Mechanisms are established to enable the team that developed or deployed AI systems to regularly incorporate adjudicated feedback from relevant AI actors into system design and implementation.
GOVERN 6: Policies and procedures are in place to address AI risks and benefits arising from third-party software and data and other supply chain issues.	**GOVERN 6.1:** Policies and procedures are in place that address AI risks associated with third-party entities, including risks of infringement of a third-party's intellectual property or other rights. **GOVERN 6.2:** Contingency processes are in place to handle failures or incidents in third-party data or AI systems deemed to be high-risk.

5.2 Map

The **MAP** function establishes the context to frame risks related to an AI system. The AI lifecycle consists of many interdependent activities involving a diverse set of actors (See Figure 3). In practice, AI actors in charge of one part of the process often do not have full visibility or control over other parts and their associated contexts. The interdependencies between these activities, and among the relevant AI actors, can make it difficult to reliably anticipate impacts of AI systems. For example, early decisions in identifying purposes and objectives of an AI system can alter its behavior and capabilities, and the dynamics of deployment setting (such as end users or impacted individuals) can shape the impacts of AI system decisions. As a result, the best intentions within one dimension of the AI lifecycle can be undermined via interactions with decisions and conditions in other, later activities.

This complexity and varying levels of visibility can introduce uncertainty into risk management practices. Anticipating, assessing, and otherwise addressing potential sources of negative risk can mitigate this uncertainty and enhance the integrity of the decision process.

The information gathered while carrying out the **MAP** function enables negative risk prevention and informs decisions for processes such as model management, as well as an initial decision about appropriateness or the need for an AI solution. Outcomes in the **MAP** function are the basis for the **MEASURE** and **MANAGE** functions. Without contextual knowledge, and awareness of risks within the identified contexts, risk management is difficult to perform. The **MAP** function is intended to enhance an organization's ability to identify risks and broader contributing factors.

Implementation of this function is enhanced by incorporating perspectives from a diverse internal team and engagement with those external to the team that developed or deployed the AI system. Engagement with external collaborators, end users, potentially impacted communities, and others may vary based on the risk level of a particular AI system, the makeup of the internal team, and organizational policies. Gathering such broad perspectives can help organizations proactively prevent negative risks and develop more trustworthy AI systems by:

- improving their capacity for understanding contexts;
- checking their assumptions about context of use;
- enabling recognition of when systems are not functional within or out of their intended context;
- identifying positive and beneficial uses of their existing AI systems;
- improving understanding of limitations in AI and ML processes;
- identifying constraints in real-world applications that may lead to negative impacts;
- identifying known and foreseeable negative impacts related to intended use of AI systems; and
- anticipating risks of the use of AI systems beyond intended use.

After completing the **MAP** function, Framework users should have sufficient contextual knowledge about AI system impacts to inform an initial go/no-go decision about whether to design, develop, or deploy an AI system. If a decision is made to proceed, organizations should utilize the **MEASURE** and **MANAGE** functions along with policies and procedures put into place in the **GOVERN** function to assist in AI risk management efforts. It is incumbent on Framework users to continue applying the **MAP** function to AI systems as context, capabilities, risks, benefits, and potential impacts evolve over time.

Practices related to mapping AI risks are described in the NIST AI RMF Playbook. Table 2 lists the **MAP** function's categories and subcategories.

Table 2: Categories and subcategories for the **MAP** function.

Categories	Subcategories
MAP 1: Context is established and understood.	**MAP 1.1:** Intended purposes, potentially beneficial uses, context-specific laws, norms and expectations, and prospective settings in which the AI system will be deployed are understood and documented. Considerations include: the specific set or types of users along with their expectations; potential positive and negative impacts of system uses to individuals, communities, organizations, society, and the planet; assumptions and related limitations about AI system purposes, uses, and risks across the development or product AI lifecycle; and related TEVV and system metrics. **MAP 1.2:** Interdisciplinary AI actors, competencies, skills, and capacities for establishing context reflect demographic diversity and broad domain and user experience expertise, and their participation is documented. Opportunities for interdisciplinary collaboration are prioritized. **MAP 1.3:** The organization's mission and relevant goals for AI technology are understood and documented. **MAP 1.4:** The business value or context of business use has been clearly defined or – in the case of assessing existing AI systems – re-evaluated. **MAP 1.5:** Organizational risk tolerances are determined and documented. **MAP 1.6:** System requirements (e.g., "the system shall respect the privacy of its users") are elicited from and understood by relevant AI actors. Design decisions take socio-technical implications into account to address AI risks.
MAP 2: Categorization of the AI system is performed.	**MAP 2.1:** The specific tasks and methods used to implement the tasks that the AI system will support are defined (e.g., classifiers, generative models, recommenders). **MAP 2.2:** Information about the AI system's knowledge limits and how system output may be utilized and overseen by humans is documented. Documentation provides sufficient information to assist relevant AI actors when making decisions and taking subsequent actions.

Continued on next page

Table 2: Categories and subcategories for the **MAP** function. (Continued)

Categories	Subcategories
	MAP 2.3: Scientific integrity and TEVV considerations are identified and documented, including those related to experimental design, data collection and selection (e.g., availability, representativeness, suitability), system trustworthiness, and construct validation.
MAP 3: AI capabilities, targeted usage, goals, and expected benefits and costs compared with appropriate benchmarks are understood.	**MAP 3.1:** Potential benefits of intended AI system functionality and performance are examined and documented. **MAP 3.2:** Potential costs, including non-monetary costs, which result from expected or realized AI errors or system functionality and trustworthiness – as connected to organizational risk tolerance – are examined and documented. **MAP 3.3:** Targeted application scope is specified and documented based on the system's capability, established context, and AI system categorization. **MAP 3.4:** Processes for operator and practitioner proficiency with AI system performance and trustworthiness – and relevant technical standards and certifications – are defined, assessed, and documented. **MAP 3.5:** Processes for human oversight are defined, assessed, and documented in accordance with organizational policies from the **GOVERN** function.
MAP 4: Risks and benefits are mapped for all components of the AI system including third-party software and data.	**MAP 4.1:** Approaches for mapping AI technology and legal risks of its components – including the use of third-party data or software – are in place, followed, and documented, as are risks of infringement of a third party's intellectual property or other rights. **MAP 4.2:** Internal risk controls for components of the AI system, including third-party AI technologies, are identified and documented.
MAP 5: Impacts to individuals, groups, communities, organizations, and society are characterized.	**MAP 5.1:** Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented.

Continued on next page

Table 2: Categories and subcategories for the MAP function. (Continued)

Categories	Subcategories
	MAP 5.2: Practices and personnel for supporting regular engagement with relevant AI actors and integrating feedback about positive, negative, and unanticipated impacts are in place and documented.

5.3 Measure

The MEASURE function employs quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts. It uses knowledge relevant to AI risks identified in the MAP function and informs the MANAGE function. AI systems should be tested before their deployment and regularly while in operation. AI risk measurements include documenting aspects of systems' functionality and trustworthiness.

Measuring AI risks includes tracking metrics for trustworthy characteristics, social impact, and human-AI configurations. Processes developed or adopted in the MEASURE function should include rigorous software testing and performance assessment methodologies with associated measures of uncertainty, comparisons to performance benchmarks, and formalized reporting and documentation of results. Processes for independent review can improve the effectiveness of testing and can mitigate internal biases and potential conflicts of interest.

Where tradeoffs among the trustworthy characteristics arise, measurement provides a traceable basis to inform management decisions. Options may include recalibration, impact mitigation, or removal of the system from design, development, production, or use, as well as a range of compensating, detective, deterrent, directive, and recovery controls.

After completing the MEASURE function, objective, repeatable, or scalable test, evaluation, verification, and validation (TEVV) processes including metrics, methods, and methodologies are in place, followed, and documented. Metrics and measurement methodologies should adhere to scientific, legal, and ethical norms and be carried out in an open and transparent process. New types of measurement, qualitative and quantitative, may need to be developed. The degree to which each measurement type provides unique and meaningful information to the assessment of AI risks should be considered. Framework users will enhance their capacity to comprehensively evaluate system trustworthiness, identify and track existing and emergent risks, and verify efficacy of the metrics. Measurement outcomes will be utilized in the MANAGE function to assist risk monitoring and response efforts. It is incumbent on Framework users to continue applying the MEASURE function to AI systems as knowledge, methodologies, risks, and impacts evolve over time.

Practices related to measuring AI risks are described in the NIST AI RMF Playbook. Table 3 lists the **MEASURE** function's categories and subcategories.

Table 3: Categories and subcategories for the **MEASURE** function.

Categories	Subcategories
MEASURE 1: Appropriate methods and metrics are identified and applied.	**MEASURE 1.1:** Approaches and metrics for measurement of AI risks enumerated during the **MAP** function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot – be measured are properly documented.
	MEASURE 1.2: Appropriateness of AI metrics and effectiveness of existing controls are regularly assessed and updated, including reports of errors and potential impacts on affected communities.
	MEASURE 1.3: Internal experts who did not serve as front-line developers for the system and/or independent assessors are involved in regular assessments and updates. Domain experts, users, AI actors external to the team that developed or deployed the AI system, and affected communities are consulted in support of assessments as necessary per organizational risk tolerance.
MEASURE 2: AI systems are evaluated for trustworthy characteristics.	**MEASURE 2.1:** Test sets, metrics, and details about the tools used during TEVV are documented.
	MEASURE 2.2: Evaluations involving human subjects meet applicable requirements (including human subject protection) and are representative of the relevant population.
	MEASURE 2.3: AI system performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for conditions similar to deployment setting(s). Measures are documented.
	MEASURE 2.4: The functionality and behavior of the AI system and its components – as identified in the **MAP** function – are monitored when in production.
	MEASURE 2.5: The AI system to be deployed is demonstrated to be valid and reliable. Limitations of the generalizability beyond the conditions under which the technology was developed are documented.

Continued on next page

Table 3: Categories and subcategories for the **MEASURE** function. (Continued)

Categories	Subcategories
	MEASURE 2.6: The AI system is evaluated regularly for safety risks – as identified in the **MAP** function. The AI system to be deployed is demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and it can fail safely, particularly if made to operate beyond its knowledge limits. Safety metrics reflect system reliability and robustness, real-time monitoring, and response times for AI system failures.
	MEASURE 2.7: AI system security and resilience – as identified in the **MAP** function – are evaluated and documented.
	MEASURE 2.8: Risks associated with transparency and accountability – as identified in the **MAP** function – are examined and documented.
	MEASURE 2.9: The AI model is explained, validated, and documented, and AI system output is interpreted within its context – as identified in the **MAP** function – to inform responsible use and governance.
	MEASURE 2.10: Privacy risk of the AI system – as identified in the **MAP** function – is examined and documented.
	MEASURE 2.11: Fairness and bias – as identified in the **MAP** function – are evaluated and results are documented.
	MEASURE 2.12: Environmental impact and sustainability of AI model training and management activities – as identified in the **MAP** function – are assessed and documented.
	MEASURE 2.13: Effectiveness of the employed TEVV metrics and processes in the **MEASURE** function are evaluated and documented.
MEASURE 3: Mechanisms for tracking identified AI risks over time are in place.	**MEASURE 3.1:** Approaches, personnel, and documentation are in place to regularly identify and track existing, unanticipated, and emergent AI risks based on factors such as intended and actual performance in deployed contexts.
	MEASURE 3.2: Risk tracking approaches are considered for settings where AI risks are difficult to assess using currently available measurement techniques or where metrics are not yet available.

Continued on next page

Table 3: Categories and subcategories for the **MEASURE** function. (Continued)

Categories	Subcategories
	MEASURE 3.3: Feedback processes for end users and impacted communities to report problems and appeal system outcomes are established and integrated into AI system evaluation metrics.
MEASURE 4: Feedback about efficacy of measurement is gathered and assessed.	**MEASURE 4.1:** Measurement approaches for identifying AI risks are connected to deployment context(s) and informed through consultation with domain experts and other end users. Approaches are documented.
	MEASURE 4.2: Measurement results regarding AI system trustworthiness in deployment context(s) and across the AI lifecycle are informed by input from domain experts and relevant AI actors to validate whether the system is performing consistently as intended. Results are documented.
	MEASURE 4.3: Measurable performance improvements or declines based on consultations with relevant AI actors, including affected communities, and field data about context-relevant risks and trustworthiness characteristics are identified and documented.

5.4 Manage

The **MANAGE** function entails allocating risk resources to mapped and measured risks on a regular basis and as defined by the **GOVERN** function. Risk treatment comprises plans to respond to, recover from, and communicate about incidents or events.

Contextual information gleaned from expert consultation and input from relevant AI actors – established in **GOVERN** and carried out in **MAP** – is utilized in this function to decrease the likelihood of system failures and negative impacts. Systematic documentation practices established in **GOVERN** and utilized in **MAP** and **MEASURE** bolster AI risk management efforts and increase transparency and accountability. Processes for assessing emergent risks are in place, along with mechanisms for continual improvement.

After completing the **MANAGE** function, plans for prioritizing risk and regular monitoring and improvement will be in place. Framework users will have enhanced capacity to manage the risks of deployed AI systems and to allocate risk management resources based on assessed and prioritized risks. It is incumbent on Framework users to continue to apply the **MANAGE** function to deployed AI systems as methods, contexts, risks, and needs or expectations from relevant AI actors evolve over time.

Practices related to managing AI risks are described in the NIST AI RMF Playbook. Table 4 lists the **MANAGE** function's categories and subcategories.

Table 4: Categories and subcategories for the **MANAGE** function.

Categories	Subcategories
MANAGE 1: AI risks based on assessments and other analytical output from the **MAP** and **MEASURE** functions are prioritized, responded to, and managed.	**MANAGE 1.1:** A determination is made as to whether the AI system achieves its intended purposes and stated objectives and whether its development or deployment should proceed. **MANAGE 1.2:** Treatment of documented AI risks is prioritized based on impact, likelihood, and available resources or methods. **MANAGE 1.3:** Responses to the AI risks deemed high priority, as identified by the **MAP** function, are developed, planned, and documented. Risk response options can include mitigating, transferring, avoiding, or accepting. **MANAGE 1.4:** Negative residual risks (defined as the sum of all unmitigated risks) to both downstream acquirers of AI systems and end users are documented.
MANAGE 2: Strategies to maximize AI benefits and minimize negative impacts are planned, prepared, implemented, documented, and informed by input from relevant AI actors.	**MANAGE 2.1:** Resources required to manage AI risks are taken into account – along with viable non-AI alternative systems, approaches, or methods – to reduce the magnitude or likelihood of potential impacts. **MANAGE 2.2:** Mechanisms are in place and applied to sustain the value of deployed AI systems. **MANAGE 2.3:** Procedures are followed to respond to and recover from a previously unknown risk when it is identified. **MANAGE 2.4:** Mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.
MANAGE 3: AI risks and benefits from third-party entities are managed.	**MANAGE 3.1:** AI risks and benefits from third-party resources are regularly monitored, and risk controls are applied and documented. **MANAGE 3.2:** Pre-trained models which are used for development are monitored as part of AI system regular monitoring and maintenance.

Continued on next page

Table 4: Categories and subcategories for the **MANAGE** function. (Continued)

Categories	Subcategories
MANAGE 4: Risk treatments, including response and recovery, and communication plans for the identified and measured AI risks are documented and monitored regularly.	**MANAGE 4.1:** Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management. **MANAGE 4.2:** Measurable activities for continual improvements are integrated into AI system updates and include regular engagement with interested parties, including relevant AI actors. **MANAGE 4.3:** Incidents and errors are communicated to relevant AI actors, including affected communities. Processes for tracking, responding to, and recovering from incidents and errors are followed and documented.

6. AI RMF Profiles

AI RMF *use-case profiles* are implementations of the AI RMF functions, categories, and subcategories for a specific setting or application based on the requirements, risk tolerance, and resources of the Framework user: for example, an AI RMF *hiring profile* or an AI RMF *fair housing profile*. Profiles may illustrate and offer insights into how risk can be managed at various stages of the AI lifecycle or in specific sector, technology, or end-use applications. AI RMF profiles assist organizations in deciding how they might best manage AI risk that is well-aligned with their goals, considers legal/regulatory requirements and best practices, and reflects risk management priorities.

AI RMF *temporal profiles* are descriptions of either the current state or the desired, target state of specific AI risk management activities within a given sector, industry, organization, or application context. An AI RMF Current Profile indicates how AI is currently being managed and the related risks in terms of current outcomes. A Target Profile indicates the outcomes needed to achieve the desired or target AI risk management goals.

Comparing Current and Target Profiles likely reveals gaps to be addressed to meet AI risk management objectives. Action plans can be developed to address these gaps to fulfill outcomes in a given category or subcategory. Prioritization of gap mitigation is driven by the user's needs and risk management processes. This risk-based approach also enables Framework users to compare their approaches with other approaches and to gauge the resources needed (e.g., staffing, funding) to achieve AI risk management goals in a cost-effective, prioritized manner.

AI RMF *cross-sectoral profiles* cover risks of models or applications that can be used across use cases or sectors. Cross-sectoral profiles can also cover how to govern, map, measure, and manage risks for activities or business processes common across sectors such as the use of large language models, cloud-based services or acquisition.

This Framework does not prescribe profile templates, allowing for flexibility in implementation.

Appendix A:
Descriptions of AI Actor Tasks from Figures 2 and 3

AI Design tasks are performed during the Application Context and Data and Input phases of the AI lifecycle in Figure 2. AI Design actors create the concept and objectives of AI systems and are responsible for the planning, design, and data collection and processing tasks of the AI system so that the AI system is lawful and fit-for-purpose. Tasks include articulating and documenting the system's concept and objectives, underlying assumptions, context, and requirements; gathering and cleaning data; and documenting the metadata and characteristics of the dataset. AI actors in this category include data scientists, domain experts, socio-cultural analysts, experts in the field of diversity, equity, inclusion, and accessibility, members of impacted communities, human factors experts (e.g., UX/UI design), governance experts, data engineers, data providers, system funders, product managers, third-party entities, evaluators, and legal and privacy governance.

AI Development tasks are performed during the AI Model phase of the lifecycle in Figure 2. AI Development actors provide the initial infrastructure of AI systems and are responsible for model building and interpretation tasks, which involve the creation, selection, calibration, training, and/or testing of models or algorithms. AI actors in this category include machine learning experts, data scientists, developers, third-party entities, legal and privacy governance experts, and experts in the socio-cultural and contextual factors associated with the deployment setting.

AI Deployment tasks are performed during the Task and Output phase of the lifecycle in Figure 2. AI Deployment actors are responsible for contextual decisions relating to how the AI system is used to assure deployment of the system into production. Related tasks include piloting the system, checking compatibility with legacy systems, ensuring regulatory compliance, managing organizational change, and evaluating user experience. AI actors in this category include system integrators, software developers, end users, operators and practitioners, evaluators, and domain experts with expertise in human factors, socio-cultural analysis, and governance.

Operation and Monitoring tasks are performed in the Application Context/Operate and Monitor phase of the lifecycle in Figure 2. These tasks are carried out by AI actors who are responsible for operating the AI system and working with others to regularly assess system output and impacts. AI actors in this category include system operators, domain experts, AI designers, users who interpret or incorporate the output of AI systems, product developers, evaluators and auditors, compliance experts, organizational management, and members of the research community.

Test, Evaluation, Verification, and Validation (TEVV) tasks are performed throughout the AI lifecycle. They are carried out by AI actors who examine the AI system or its components, or detect and remediate problems. Ideally, AI actors carrying out verification

and validation tasks are distinct from those who perform test and evaluation actions. Tasks can be incorporated into a phase as early as design, where tests are planned in accordance with the design requirement.

- TEVV tasks for design, planning, and data may center on internal and external validation of assumptions for system design, data collection, and measurements relative to the intended context of deployment or application.
- TEVV tasks for development (i.e., model building) include model validation and assessment.
- TEVV tasks for deployment include system validation and integration in production, with testing, and recalibration for systems and process integration, user experience, and compliance with existing legal, regulatory, and ethical specifications.
- TEVV tasks for operations involve ongoing monitoring for periodic updates, testing, and subject matter expert (SME) recalibration of models, the tracking of incidents or errors reported and their management, the detection of emergent properties and related impacts, and processes for redress and response.

Human Factors tasks and activities are found throughout the dimensions of the AI lifecycle. They include human-centered design practices and methodologies, promoting the active involvement of end users and other interested parties and relevant AI actors, incorporating context-specific norms and values in system design, evaluating and adapting end user experiences, and broad integration of humans and human dynamics in all phases of the AI lifecycle. Human factors professionals provide multidisciplinary skills and perspectives to understand context of use, inform interdisciplinary and demographic diversity, engage in consultative processes, design and evaluate user experience, perform human-centered evaluation and testing, and inform impact assessments.

Domain Expert tasks involve input from multidisciplinary practitioners or scholars who provide knowledge or expertise in – and about – an industry sector, economic sector, context, or application area where an AI system is being used. AI actors who are domain experts can provide essential guidance for AI system design and development, and interpret outputs in support of work performed by TEVV and AI impact assessment teams.

AI Impact Assessment tasks include assessing and evaluating requirements for AI system accountability, combating harmful bias, examining impacts of AI systems, product safety, liability, and security, among others. AI actors such as impact assessors and evaluators provide technical, human factor, socio-cultural, and legal expertise.

Procurement tasks are conducted by AI actors with financial, legal, or policy management authority for acquisition of AI models, products, or services from a third-party developer, vendor, or contractor.

Governance and Oversight tasks are assumed by AI actors with management, fiduciary, and legal authority and responsibility for the organization in which an AI system is de-

signed, developed, and/or deployed. Key AI actors responsible for AI governance include organizational management, senior leadership, and the Board of Directors. These actors are parties that are concerned with the impact and sustainability of the organization as a whole.

Additional AI Actors

Third-party entities include providers, developers, vendors, and evaluators of data, algorithms, models, and/or systems and related services for another organization or the organization's customers or clients. Third-party entities are responsible for AI design and development tasks, in whole or in part. By definition, they are external to the design, development, or deployment team of the organization that acquires its technologies or services. The technologies acquired from third-party entities may be complex or opaque, and risk tolerances may not align with the deploying or operating organization.

End users of an AI system are the individuals or groups that use the system for specific purposes. These individuals or groups interact with an AI system in a specific context. End users can range in competency from AI experts to first-time technology end users.

Affected individuals/communities encompass all individuals, groups, communities, or organizations directly or indirectly affected by AI systems or decisions based on the output of AI systems. These individuals do not necessarily interact with the deployed system or application.

Other AI actors may provide formal or quasi-formal norms or guidance for specifying and managing AI risks. They can include **trade associations, standards developing organizations, advocacy groups, researchers, environmental groups, and civil society organizations**.

The general public is most likely to directly experience positive and negative impacts of AI technologies. They may provide the motivation for actions taken by the AI actors. This group can include individuals, communities, and consumers associated with the context in which an AI system is developed or deployed.

Appendix B:
How AI Risks Differ from Traditional Software Risks

As with traditional software, risks from AI-based technology can be bigger than an enterprise, span organizations, and lead to societal impacts. AI systems also bring a set of risks that are not comprehensively addressed by current risk frameworks and approaches. Some AI system features that present risks also can be beneficial. For example, pre-trained models and transfer learning can advance research and increase accuracy and resilience when compared to other models and approaches. Identifying contextual factors in the **MAP** function will assist AI actors in determining the level of risk and potential management efforts.

Compared to traditional software, AI-specific risks that are new or increased include the following:

- The data used for building an AI system may not be a true or appropriate representation of the context or intended use of the AI system, and the ground truth may either not exist or not be available. Additionally, harmful bias and other data quality issues can affect AI system trustworthiness, which could lead to negative impacts.
- AI system dependency and reliance on data for training tasks, combined with increased volume and complexity typically associated with such data.
- Intentional or unintentional changes during training may fundamentally alter AI system performance.
- Datasets used to train AI systems may become detached from their original and intended context or may become stale or outdated relative to deployment context.
- AI system scale and complexity (many systems contain billions or even trillions of decision points) housed within more traditional software applications.
- Use of pre-trained models that can advance research and improve performance can also increase levels of statistical uncertainty and cause issues with bias management, scientific validity, and reproducibility.
- Higher degree of difficulty in predicting failure modes for emergent properties of large-scale pre-trained models.
- Privacy risk due to enhanced data aggregation capability for AI systems.
- AI systems may require more frequent maintenance and triggers for conducting corrective maintenance due to data, model, or concept drift.
- Increased opacity and concerns about reproducibility.
- Underdeveloped software testing standards and inability to document AI-based practices to the standard expected of traditionally engineered software for all but the simplest of cases.
- Difficulty in performing regular AI-based software testing, or determining what to test, since AI systems are not subject to the same controls as traditional code development.

- Computational costs for developing AI systems and their impact on the environment and planet.
- Inability to predict or detect the side effects of AI-based systems beyond statistical measures.

Privacy and cybersecurity risk management considerations and approaches are applicable in the design, development, deployment, evaluation, and use of AI systems. Privacy and cybersecurity risks are also considered as part of broader enterprise risk management considerations, which may incorporate AI risks. As part of the effort to address AI trustworthiness characteristics such as "Secure and Resilient" and "Privacy-Enhanced," organizations may consider leveraging available standards and guidance that provide broad guidance to organizations to reduce security and privacy risks, such as, but not limited to, the NIST Cybersecurity Framework, the NIST Privacy Framework, the NIST Risk Management Framework, and the Secure Software Development Framework. These frameworks have some features in common with the AI RMF. Like most risk management approaches, they are outcome-based rather than prescriptive and are often structured around a Core set of functions, categories, and subcategories. While there are significant differences between these frameworks based on the domain addressed – and because AI risk management calls for addressing many other types of risks – frameworks like those mentioned above may inform security and privacy considerations in the **MAP**, **MEASURE**, and **MANAGE** functions of the AI RMF.

At the same time, guidance available before publication of this AI RMF does not comprehensively address many AI system risks. For example, existing frameworks and guidance are unable to:

- adequately manage the problem of harmful bias in AI systems;
- confront the challenging risks related to generative AI;
- comprehensively address security concerns related to evasion, model extraction, membership inference, availability, or other machine learning attacks;
- account for the complex attack surface of AI systems or other security abuses enabled by AI systems; and
- consider risks associated with third-party AI technologies, transfer learning, and off-label use where AI systems may be trained for decision-making outside an organization's security controls or trained in one domain and then "fine-tuned" for another.

Both AI and traditional software technologies and systems are subject to rapid innovation. Technology advances should be monitored and deployed to take advantage of those developments and work towards a future of AI that is both trustworthy and responsible.

Appendix C:
AI Risk Management and Human-AI Interaction

Organizations that design, develop, or deploy AI systems for use in operational settings may enhance their AI risk management by understanding current limitations of human-AI interaction. The AI RMF provides opportunities to clearly define and differentiate the various human roles and responsibilities when using, interacting with, or managing AI systems.

Many of the data-driven approaches that AI systems rely on attempt to convert or represent individual and social observational and decision-making practices into measurable quantities. Representing complex human phenomena with mathematical models can come at the cost of removing necessary context. This loss of context may in turn make it difficult to understand individual and societal impacts that are key to AI risk management efforts.

Issues that merit further consideration and research include:

1. **Human roles and responsibilities in decision making and overseeing AI systems need to be clearly defined and differentiated.** Human-AI configurations can span from fully autonomous to fully manual. AI systems can autonomously make decisions, defer decision making to a human expert, or be used by a human decision maker as an additional opinion. Some AI systems may not require human oversight, such as models used to improve video compression. Other systems may specifically require human oversight.

2. **Decisions that go into the design, development, deployment, evaluation, and use of AI systems reflect systemic and human cognitive biases.** AI actors bring their cognitive biases, both individual and group, into the process. Biases can stem from end-user decision-making tasks and be introduced across the AI lifecycle via human assumptions, expectations, and decisions during design and modeling tasks. These biases, which are not necessarily always harmful, may be exacerbated by AI system opacity and the resulting lack of transparency. Systemic biases at the organizational level can influence how teams are structured and who controls the decision-making processes throughout the AI lifecycle. These biases can also influence downstream decisions by end users, decision makers, and policy makers and may lead to negative impacts.

3. **Human-AI interaction results vary.** Under certain conditions – for example, in perceptual-based judgment tasks – the AI part of the human-AI interaction can amplify human biases, leading to more biased decisions than the AI or human alone. When these variations are judiciously taken into account in organizing human-AI teams, however, they can result in complementarity and improved overall performance.

4. **Presenting AI system information to humans is complex.** Humans perceive and derive meaning from AI system output and explanations in different ways, reflecting different individual preferences, traits, and skills.

The **GOVERN** function provides organizations with the opportunity to clarify and define the roles and responsibilities for the humans in the Human-AI team configurations and those who are overseeing the AI system performance. The **GOVERN** function also creates mechanisms for organizations to make their decision-making processes more explicit, to help counter systemic biases.

The **MAP** function suggests opportunities to define and document processes for operator and practitioner proficiency with AI system performance and trustworthiness concepts, and to define relevant technical standards and certifications. Implementing **MAP** function categories and subcategories may help organizations improve their internal competency for analyzing context, identifying procedural and system limitations, exploring and examining impacts of AI-based systems in the real world, and evaluating decision-making processes throughout the AI lifecycle.

The **GOVERN** and **MAP** functions describe the importance of interdisciplinarity and demographically diverse teams and utilizing feedback from potentially impacted individuals and communities. AI actors called out in the AI RMF who perform human factors tasks and activities can assist technical teams by anchoring in design and development practices to user intentions and representatives of the broader AI community, and societal values. These actors further help to incorporate context-specific norms and values in system design and evaluate end user experiences – in conjunction with AI systems.

AI risk management approaches for human-AI configurations will be augmented by ongoing research and evaluation. For example, the degree to which humans are empowered and incentivized to challenge AI system output requires further studies. Data about the frequency and rationale with which humans overrule AI system output in deployed systems may be useful to collect and analyze.

Appendix D:
Attributes of the AI RMF

NIST described several key attributes of the AI RMF when work on the Framework first began. These attributes have remained intact and were used to guide the AI RMF's development. They are provided here as a reference.

The AI RMF strives to:

1. Be risk-based, resource-efficient, pro-innovation, and voluntary.
2. Be consensus-driven and developed and regularly updated through an open, transparent process. All stakeholders should have the opportunity to contribute to the AI RMF's development.
3. Use clear and plain language that is understandable by a broad audience, including senior executives, government officials, non-governmental organization leadership, and those who are not AI professionals – while still of sufficient technical depth to be useful to practitioners. The AI RMF should allow for communication of AI risks across an organization, between organizations, with customers, and to the public at large.
4. Provide common language and understanding to manage AI risks. The AI RMF should offer taxonomy, terminology, definitions, metrics, and characterizations for AI risk.
5. Be easily usable and fit well with other aspects of risk management. Use of the Framework should be intuitive and readily adaptable as part of an organization's broader risk management strategy and processes. It should be consistent or aligned with other approaches to managing AI risks.
6. Be useful to a wide range of perspectives, sectors, and technology domains. The AI RMF should be universally applicable to any AI technology and to context-specific use cases.
7. Be outcome-focused and non-prescriptive. The Framework should provide a catalog of outcomes and approaches rather than prescribe one-size-fits-all requirements.
8. Take advantage of and foster greater awareness of existing standards, guidelines, best practices, methodologies, and tools for managing AI risks – as well as illustrate the need for additional, improved resources.
9. Be law- and regulation-agnostic. The Framework should support organizations' abilities to operate under applicable domestic and international legal or regulatory regimes.
10. Be a living document. The AI RMF should be readily updated as technology, understanding, and approaches to AI trustworthiness and uses of AI change and as stakeholders learn from implementing AI risk management generally and this framework in particular.

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.100-1

Technology Letter 25-01

February 2025

Subject:

Information Technology (IT) Policies for Generative Artificial Intelligence (GenAI)

Reference:

Executive Order (EO): N-12-23

Assembly Bill (AB): 2013, 2885

Senate Bill (SB): 896

Government Code (GC): 11000, 11545, 11546, 11546.1(e), 11546.7, 11549.63, 11549.63(g), 11549.64, 11549.65, 11549.65(c), 11549.66

State Administrative Manual (SAM): 4819.2, 4819.35, 4940, 4983.1, 4986.1, 4986.2, 4986.3, 4986.4, 4986.5, 4986.6, 4986.7, 4986.8, 4986.9, 4986.10, 4986.11, 4986.12, 4986.13, 5210, 5335

State Contracting Manual (SCM): SCM, Purchasing Authority Standards, 100.3

State Information Management Manual (SIMM): 18B, 19H, 22, 45, 50, 71A 71B 140, 141, 5305-F, 5310-C, 5325-A; 5325-B, 5335-A, 5340-A, 5340-C, 5345-A

California Civil Code:1798.29, 1798.3

National Institute of Standards and Technology (NIST) Artificial Intelligence (AI) Risk Management Framework (RMF), 800-53

Background

The California Executive Order N-12-23 (GenAI EO) signed by Governor Newsom on September 6, 2023, guides the state to prepare for the advancements and potential risks associated with Generative Artificial Intelligence (GenAI). Senate Bill (SB) 896 known as the GenAI Accountability Act aims to regulate and mitigate risks associated with GenAI technologies in California by requiring transparency, risk assessments, and clear communication when Artificial Intelligence (AI) is used in state government services. In addition, Assembly Bill (AB) 2013 mandates the public disclosure of the data used to train GenAI models, including summaries of the data sets and their sources. Lastly, AB 2885 establishes uniform definitions for AI in California law. Together with the GenAI EO, these new laws emphasize California's commitment to leading the world in ethical, transparent, and responsible GenAI development and outline measures to ensure the safe and equitable deployment of GenAI.

The Department of Technology (CDT) recognizes the tremendous potential of AI and GenAI to improve the lives of California residents, support the state work force, and improve the efficiency of services delivered by the state. It also recognizes the use of AI and GenAI must be guided by principles of fairness, transparency, privacy, security, and accountability to ensure that the systems and technology used are protected.

The GenAI policies described below outline the requirements for the responsible and secure use of GenAI within the State of California. These GenAI policies aim to ensure efficiency, effectiveness, accessibility, confidentiality, integrity, security, and privacy of State data. These policies apply to all GenAI procured or developed by a state entity. Except as otherwise indicated in the specific GenAI policy, these GenAI policies apply to all State Entities, employees, contractors, and authorized users who access, use, or manage GenAI within the State, as applicable. Additionally, processes, procedures, forms, and contract language noted below have been created, revised or updated to align with the GenAI EO, legislation, and policies.

Purpose:

The purpose of this Technology Letter (TL) is to announce the following new and revised policies, processes, procedures, forms, and contract language for GenAI that are effective as of February 20, 2025:

- **SAM 4819.2 Definitions:** Updates the state Information Technology (IT)

definitions based on new California legislation, AB 2885.

- **SAM 4986.1 GenAI Policy Introduction:** Sets forth the requirements for the responsible and secure use of GenAI within the State. The policy focuses on minimizing risks while ensuring efficiency, confidentiality, and integrity to processes and operations.
- **SAM 4986.2 Definitions for GenAI:** Provides the definitions specific to the GenAI policy.
- **SAM 4986.3 GenAI Use Identification and High-Risk Inventory:** Outlines the requirements to identify, track, and maintain an inventory of high-risk GenAI for annual reporting to the State legislature.
- **SAM 4986.4 GenAI Training Data and Transparency:** Requires transparency about the data used to train GenAI systems and services by posting relevant documentation on the state entity's websites by January 1, 2026.
- **SAM 4986.5 GenAI Hosting:** Mandates that any new, expanded, or refreshed GenAI must adhere to the Cloud Computing Policy, SAM 4983.1.
- **SAM 4986.6 GenAI Proof of Concept (POC) and Minimum Viable Product (MVP):** Mandates that all GenAI POCs be hosted in a CDT "sandbox" environment and all MVPs in a CDT-approved production environment. POCs and MVPs must be tested for feasibility, performance validation, and risk mitigation strategies during these phases.
- **SAM 4986.7 IT Projects Utilizing GenAI Planning and Approval:** Mandates specific assessments and project planning prior to procuring or developing an IT Project Utilizing GenAI.
- **SAM 4986.8 IT Projects Utilizing GenAI Project Management:** Requires any State Entity approved to proceed with an IT Project utilizing GenAI to follow the SIMM 19H Project Delivery Lifecycle (PDL) to ensure proper management and oversight.
- **SAM 4986.9 GenAI Procurement:** Requires compliance with GenAI procurement policies and disclosures as outlined below.

        - SCM, Volume 2, Chapter 23, including but not limited to, GenAI disclosure language for solicitations, updated ITGP (Cloud) and (Non-Cloud) with GenAI terms and conditions, new GenAI Special Provisions as needed, and
        - SAM 4986.9 including but not limited to compliance with security and privacy standards for GenAI and completion of SIMM 5305-F GenAI Risk Assessment.
        - All solicitations and contracts for IT and telecommunications goods and services require disclosure notification that applies to GenAI as a contract deliverable or when GenAI has a material impact on risk, functionality, or contract performance.

- **SAM 4986.10 Privacy for GenAI:** Requires State Entities using or procuring GenAI to protect data privacy and security according to state and federal laws and policies.
- **SAM 4986.11 Security for GenAI:** Outlines security incident reporting and compliance requirements for GenAI, including the State Entity's Incident Response Plan, and prompt reporting of violations to CDT.
- **SAM 4986.12 Acceptable Use of GenAI:** Requires education and compliance with the Acceptable Use Policy. Prohibits the use of GenAI for illicit, controversial, or biased content. Outlines acceptable use of GenAI.
- **SAM 4986.13 GenAI Workforce Training:** Requires state personnel to complete the necessary GenAI training relevant to their roles to ensure the proper understanding and use of GenAI.
- **SIMM 19H Project Delivery Lifecycle (PDL):** Introduces a flexible, innovative, and responsive approach that adapts to rapidly evolving business needs and technologies.
- **SIMM 71A Certification of Compliance with IT Policies Instructions:** The instructions have been updated to include GenAI requirements.

- **SIMM 71B Certification of Compliance with IT Policies Template:** The template has been updated to include GenAI requirements.
- **SIMM 5305-F GenAI Risk Assessment:** Updates include format changes, revised questions, and risk assessment tables. There are two new sections to the document: safeguards and data types.
- **Form STD 1000 GenAI Reporting and Factsheet:** This statewide form has been retired and is no longer required.

Questions:

Direct questions regarding this Technology Letter to the Department of Technology, State IT Policy Office at ITpolicy@state.ca.gov.

Signature:

On file

Liana Bailey-Crimmins, State Chief Information Officer and Director

California Department of Technology

Website Accessibility Certification

San Francisco Generative AI Guidelines (2025)

For All City and County personnel, including employees, contractors, consultants, volunteers, and vendors working on behalf of the City

July 7th, 2025

Top 5 Guidelines for Using Generative AI

1. Whether generated by AI or a human, you are ultimately responsible for any content you use or share.
2. Copilot Chat is available to City employees as a secure option for most Generative AI tasks. Other secure enterprise tools may also be available to staff. We strongly discourage the use of Generative AI tools that have not been purchased or vetted by the City. If you must use public or consumer Generative AI tools, never enter sensitive, confidential, or City data, as these tools may store or reuse the information you provide.
3. Always review, edit, fact-check, validate, and/or test AI generated content, which isn't always accurate.
4. Always disclose AI use when it contributes to public-facing or sensitive work or required by regulations. Ensure the Generative AI tool is properly recorded in the City's 22J inventory and provide direct notice to impacted individuals.
5. Never use Generative AI to generate deepfakes—fake images or recordings--or other content that could be mistakenly interpreted by someone to be real.

1. Introduction

Enterprise Generative AI (GenAI) tools procured and licensed by the Department of Technology (DT) are now available for use by staff of the City and County of San Francisco (City), opening new opportunities to improve the effectiveness, efficiency, and responsiveness of City services for all San Franciscans.

Unlike other AI technologies used by the City—which support informed decisions based on input data—GenAI tools can generate new content based on patterns identified in large datasets, often in a matter of seconds. Examples include text, images, music, and code.

However, the use of GenAI technology by the City also poses unique risks to its workers, residents, and visitors. While AI-generated content can appear authoritative and polished, it can be inaccurate, biased, or misleading. GenAI use can also heighten the risk of privacy breaches, unauthorized data sharing, and cybersecurity threats. Furthermore, overreliance on GenAI for decisions that affect the public's rights or safety can reduce transparency, weaken accountability, and erode trust in government.

1.1. Purpose of the Guidelines

These guidelines are designed to help City staff use GenAI tools effectively and responsibly, while maintaining public trust, protecting resident data, and preserving the integrity of City systems.

Although GenAI tools provided by DT offer stronger privacy and security protections compared to public or consumer AI tools, they still pose significant risks. These include bias, misinformation, factual errors, hallucinations, copyright violations, and inconsistent outputs. Such risks are heightened when City employees rely on these tools without exercising the necessary human oversight. To mitigate these risks, all GenAI outputs should be carefully reviewed before use to ensure accuracy, fairness, and alignment with City values and policies. City staff must also adhere to existing data protection and governance requirements, and provide transparency to the public about when and how GenAI tools are being used, in compliance with the City AI Transparency Ordinance (Chapter 22J).

1.2. Scope of the Guidelines

With some important distinctions, **the guidelines outlined in Sections 2.1 through 2.3 of this document apply to both:**

- **Enterprise GenAI tools**—like ChatGPT Enterprise, Microsoft Copilot, Snowflake Cortex, Adobe apps and Express, and other approved systems—licensed and managed through the Department of Technology (DT). These tools:
    - Have been **procured and configured** for City use;
    - **Allow use of sensitive City data (and restrict any vendor use of City data for AI training).** This includes protected health information (PHI) and personally identifiable information (PII) when using tools that have Business Associate Agreements (BAAs) in place, such as Microsoft Copilot and Snowflake Cortex;
    - Have **passed cybersecurity risk assessments;** and
    - Support the City's data retention standards and obligations for public request for access

- **Public or consumer GenAI tools** — such as free online chatbots or apps not procured or managed by the City. These tools do not provide adequate privacy protections, and information shared with them is often used to train the underlying models, posing potential confidentiality risks.

While the use of public or consumer GenAI tools for City business is strongly discouraged, we recognize that such tools may still be used in limited circumstances. **If you want to use public or consumer GenAI tools, you must obtain prior departmental approval and follow the precautions outlined in Section 2.4.**

2. City Guidelines for Generative AI Use

The GenAI uses outlined below are grouped by risk level, each with corresponding mitigation strategies and disclosure requirements.

As a general rule, City employees must always thoroughly review, edit, fact-check, validate, and/or test their output, as applicable. **You are ultimately responsible for any content you use or share.**

2.1. Low-Risk Use: Internal Efficiency Tasks Performed Using Enterprise Generative AI Tools

You may use City-procured AI tools for:

- Drafting internal emails, memos, or communications
- Creating summaries of meetings, documents, or reports
- Writing, editing, or debugging code
- Generating outlines or first drafts of internal materials
- Improving language access between the general public and City staff.

These uses help improve efficiency and reduce workload, but you remain the expert reviewer.

Safeguards and Responsibilities

Even when handling simple tasks, AI can make errors, omit context, or return outdated or biased information. There's no such thing as zero risk. To ensure responsible and safe use of AI tools:

- For summaries or memos, only use material you are already familiar with so you can spot issues. You should be able to independently evaluate the quality and correctness of the output.
- Always double-check factual claims, hyperlinks, and references to ensure content is backed by evidence.
- Always review and edit AI's output critically before using or sharing it.
- Only use GenAI for coding tasks if you already know the programming language. You must be able to review, debug, and test any AI-generated code.

Disclosure Requirements

Disclosure is not required when using AI for internal drafting or support. You remain fully responsible for any content you use or share, including any errors or omissions introduced by AI.

2.2. Medium- to High-Risk Use: Public-Facing or Sensitive Work Performed Using Enterprise Generative AI Tools

Use extra caution and follow additional steps when City-approved AI tools are used to perform tasks affecting public communication, services, or decisions such as:

- Drafting or translating public-facing content.
- Drafting interview questions and screening materials for hiring processes.
- Summarizing policy-related data.
- Supporting decisions related to services, enforcement, or eligibility.
- Contributing to documents that affect regulation or safety.

For these use-cases, AI can serve as a support tool, but it should never make final decisions that affect individuals or public outcomes.

Safeguards and Responsibilities

To ensure responsible and safe use of Enterprise AI tools used to perform public-facing or sensitive work:

- Only use GenAI if you have deep subject-matter expertise to review its output. This ensures you can spot errors, detect harmful implications, review suggestions critically, and make informed decisions based in AI-generated content.
- Review and edit AI-generated outputs to ensure they reflect the City's values, promote equity, and uphold ethical standards and digital accessibility standards including the use of inclusive language and design practices that serve residents with visual, cognitive, or motor impairments.
- Actively monitor for instances of bias and correct them manually.

Disclosure Requirements

When using Enterprise GenAI tools for public-facing or sensitive work, usage should be properly documented through the 22J process.

To promote transparency toward the public, individuals impacted by AI systems must receive a direct notice disclosing that GenAI substantially contributed to the work product. At minimum, this direct notice must include: a clear plain-language, multilingual statement that GenAI was used; the name and version of the tool used; a note indicating that the content was reviewed by City staff; and contact information for questions, appeals, or corrections.

> ***Direct Notice – Sample Template (please check with your department for specific requirements or approved language)***
>
> *This content was generated with the assistance of [Tool Name, Version] and reviewed by City staff for accuracy.*
> *For questions, appeals, or corrections, contact: [Email].*
>
> *View this notice in: [Español] · [中文] · [Tagalog]*
> *Or contact [email] for additional options.*

Depending on the type of content and its intended audience, other regulatory requirements related to disclosures or disclaimers, such as those under AB 3030, may also apply. Staff should always consult with their department for specific guidance on compliance with applicable laws and policies.

Furthermore, if you **paraphrase, quote, or incorporate any AI-generated content** into your own work (whether text, image, data, or other), you should **cite it appropriately, j**ust as you would any external source. Established citation guidelines, such as those from the MLA Style Center, recommend including the following details: a description of the **prompt; t**he **name and version** of the AI tool; the **developer/company; t**he **date of interaction; and a URL** to the tool or session, if applicable.

Example: Quoting AI-Generated Text

When asked to describe the symbolism of the green light in The Great Gatsby, ChatGPT provided a summary about optimism, the unattainability of the American dream, greed, and covetousness. However, when further prompted to cite the source on which that summary was based, it noted that it lacked "the ability to conduct research or cite sources independently" but that it could "provide a list of scholarly sources related to the symbolism of the green light in The Great Gatsby" ("In 200 words").

Works-Cited-List Entry (MLA Style):

"In 200 words, describe the symbolism of the green light in *The Great Gatsby*" follow-up prompt to list sources. *ChatGPT*, 13 Feb. version, OpenAI, 9 Mar. 2023, https://chat.openai.com/chat.

If a GenAI tool references or summarizes secondary sources (e.g., articles, studies, or reports), you must verify the original material and cite the primary source directly, just as you would in traditional research.

2.3. Prohibited Uses

To protect public trust, safety, and ethical standards, do **not** use GenAI tools for any of the following:

- Relying on AI to create City official documents or make decisions without expert human review.
- Generating images, audio, or video that could be mistaken for real people (including public officials or members of the public).
- Creating "deepfakes" or impersonations of any person or official—even with disclaimers.
- Fabricating fictional survey respondents or public input for research or outreach purposes.
- Relying on AI to review legal or regulatory issues.

2.4. Using Public or Consumer Generative AI Tools: Additional Guidance

The use of public or consumer Generative AI tools in place of City-approved enterprise solutions is strongly discouraged. Any experimental use of non-approved Generative AI technologies must receive prior departmental approval.

To ensure the security of City systems and data and best serve the public when using public or consumer GenAI tools, follow this guidance in addition to the ones outlined for Enterprise GenAI Tools:

- **Never enter sensitive or protected data that cannot be fully released to the public, including personal information, health information, City data, and/or financial information.** This information can be viewed by the companies that make the tools and, in some cases, other members of the public. Once entered, this information becomes part of the public record. The handling and disclosure of sensitive information is already governed by several City policies, including but not limited to:

- Charter Section 16.130, Privacy First Policy
- Administrative Code Section 12M.2(a), Nondisclosure of Private Information
- Campaign & Governmental Conduct Code section 3.228, Disclosure or Use of Confidential City Information
- The "Computers and Data Information Systems" section of the Department of Human Resource's Employee Handbook (January 2012, page 48)
- Please refer to Citywide Data Classification Standard for more specifics on data classification and department responsible roles

- **Never conceal the use of public or consumer GenAI tools** from your coworkers and remain transparent about when and how these tools are used in your work.

2.5. Guidance for Departmental IT Leaders

Departmental IT leaders have a responsibility to support right-sized GenAI uses that deliver the greatest public benefit. This begins with ensuring that AI projects are problem-led and not technology-led, by centering the specific needs and challenges faced by the department and the communities it services.

In parallel, Departmental IT leaders must ensure responsibility and transparency in how GenAI tools are used, especially when they could influence department decisions, erode the public's rights or safety, or affect access to critical services. This responsibility includes ensuring that their department complies with the transparency requirements set forth in Chapter 22J of the Administrative Code.

To support these goals, IT leaders should adhere to the following guidance:

- If your department is considering adopting or piloting a new GenAI tool, please reach out to the Emerging Technology Team at ai@sfgov.org early in the process to ensure alignment with citywide standards, policy requirements, and support.
- When purchasing technology that includes GenAI, collect the information required by 22J from vendors during contract execution or shortly thereafter. This includes the name of the technology and vendor; a description of its purpose, function, intended use, and operational context; training and generated data; an explanation of how the technology works; what it is optimizing for and its accuracy (preferably with numerical metrics); conditions affecting performance; testing for bias (including results); procedures for reporting issues or incidents; oversight mechanisms; and whether collected data may be used to train proprietary or third-party systems.
- Designate at least one 22J Lead responsible for:
  - Managing compliance with 22J.
  - Coordinating inventory submissions and acting as the point of contact for the Emerging Technology Team.
  - Determining whether each GenAI technology your department uses--or is planning to procure, borrow, or receive--qualifies for an exemption under 22J.
  - Submitting required information for non-exempt GenAI tools under 22J through LogicGate 22J or an MS form.

    - Notifying DT of any updates if GenAI tools are decommissioned, replaced or added to your department's inventory.
    - Supporting annual compliance reviews by the City Controller.
- Encourage staff to participate in GenAI–related training offered by the City to understand risks, use cases, and functionalities of specific Enterprise GenAI tools.
- If a Gen-AI feature becomes available to staff without your department's prior approval, contact the vendor to learn about privacy and security implications, and deactivate the functionality if needed. If the tool is licensed and managed by DT, report the issue immediately to the Emerging Technology Team at ai@sfgov.org.

3. Data Protection Requirements

The use of City data in Enterprise AI tools is subject to the following restrictions:

- For **Copilot Chat** and **Snowflake**, you can use **Level 4 data and below** (Levels 1–4).
- For **ChatGPT Enterprise**, you can use **Level 3 data and below** (Levels 1–3).
- **Only use PHI (Protected Health Information) in tools that have a BAA (Business Associate Agreement) in place, such as Copilot Chat and Snowflake subject to your department's approval**.
- To verify which types of department-specific data are permitted for use with City-approved tools, always check with your Department.
- **Do not enter any sensitive or protected data including personal information, health information, and/or financial information into public or consumer AI tools** not provisioned or approved for City use.

Content entered into or generated by GenAI tools may constitute public records and may be subject to disclosure under the **California Public Records Act (CPRA)**, as well as the **City and County of San Francisco's public access and records retention requirements**, including the **Sunshine Ordinance**. For Copilot Chat, user inputs and AI-generated content are retained for 30 days, consistent with the retention period for Microsoft Teams. For ChatGPT Enterprise, content is retained for 90 days.

4. Data Governance

Generative AI tools depend on accurate, clean, and well-organized data. If City data is outdated, inconsistent, or poorly maintained, AI outputs may be inaccurate or misleading.

Every department has a role in:

- Keeping records and metadata up to date
- Using standardized formats per the City's Data Management Policy
- Ensuring data is managed according to City policies

Strong data governance supports reliable and fair AI usage.

5. Background

5.1. Current Legislative and Regulatory Landscape

Since the release of the City's first GenAI guidelines in December 2023, the national regulatory landscape has shifted significantly. President Biden's 2023 Executive Order, which prioritized AI safety, security, and responsible use, was rescinded in early 2025 and replaced with a lighter-touch, pro-innovation policy that emphasizes deregulation and national competitiveness over public safeguards. At the state level, Governor Newsom reaffirmed California's leadership in responsible AI governance with the release of a *Report on Frontier AI Policy* in June 2025. The report highlights growing risks--like AI-generated scams, disinformation, and threats involving dangerous materials--and calls for strong, practical safeguards. These include enhanced transparency, independent evaluations, whistleblower protections, and systems to track harmful incidents. Built on a "trust but verify" approach, the report argues that thoughtful regulation supports, rather than hinders, responsible innovation.5.2. San Francisco: the Home of AI

San Francisco is at the center of the AI boom, with many of the world's most influential AI companies--including OpenAI, Anthropic, Databricks--as well as a growing ecosystem of startups and research labs headquartered in the City. As the home of AI, San Francisco is also on the front lines of addressing its real-world impacts on residents, workers, and public services. Given the scale of innovation underway, and the potential consequences of inaction, the City has both a responsibility and an opportunity to lead in the civic use of AI through responsible and accountable governance.

While the City will continue to track federal and state developments, it is also charting its own course under Mayor Lurie's leadership by embracing intentional, practical AI adoption grounded in ethical standards and public trust—one that delivers more effective, efficient, and responsive services to all San Franciscans.

5.3. Development of Guidelines

The Emerging Technology Team developed these updated GenAI-focused guidelines in close coordination with the City's AI Advisory Committee, a staff working group that provides guidance on the adoption, governance, and ethical use of emerging technologies in the City.

6. Contact & Versioning

The AI Advisory Committee will regularly update the City Guidelines for Generative AI Use to reflect new law, regulations, lessons learned from application, and developments in GenAI technology. Check these Guidelines regularly for updates, and bookmark to stay informed.

For questions or help with tool selection, training opportunities, or policy interpretation please check with the Emerging Technology Team at ai@sfgov.org.

This report includes content drafted with support from ChatGPT Enterprise (GPT-4o, OpenAI, June 2025). All content was reviewed and finalized by the Emerging Technology Team.

Glossary

Algorithms: are a set of rules that a machine follows to generate an outcome or a decision.

Artificial Intelligence (AI): refers to a group of technologies that can perform complex cognitive tasks like recognizing and classifying images or powering autonomous vehicles. Many AI systems are built using machine learning models. For a task like image recognition, the model learns pixel patterns from a large dataset of existing images and uses these patterns to recognize and classify new images.

Auditability for AI: AI where the outputs are explainable, monitored and validated on a regular basis.

Bard: is a conversational Gen AI chatbot built by Google

Black box models: are those where you cannot effectively determine how or why a model produced a specific result.

Chatbots: are computer programs that simulate conversations. Chatbots have been around for a few decades. Basic chatbots (without Gen AI) use ML to understand human prompts and provide more-or-less scripted answers that can guide users through a process. Gen AI chatbots can provide more human-like, conversational answers.

ChatGPT: is a conversational Gen AI chatbot built by OpenAI

Dall-e: is a Gen AI application that can generate images based on text prompts

Discriminative AI: In contrast to Gen AI, Discriminative AI models do not generate new content but can be used to predict quantities (for example, predicting home prices) or to assign group membership (for example, classifying images).

Enterprise Generative AI Tool: City-approved GenAI tool procured and managed by the Department of Technology or another City department. These tools are configured for City use, support sensitive data (with BAAs where applicable), meet cybersecurity standards, and follow strict privacy, legal, and data retention requirements.

Generative AI (Gen AI): refers to a group of technologies that can generate new content based on a user provided prompt. Many are powered by LLMs.

Large language models (LLMs): are a type of machine learning model trained using large amounts of text data. These models learn nuanced patterns and structure of language. This allows the model to understand a user generated prompt and provide a text response that is coherent. The responses are based on predicting the most likely word in a sequence of words and as a result, the answers are not always contextually correct. The training datasets used to build these models can contain gender, racial,

political and other biases. Since the models have learnt from biased data, their outputs can reflect these biases. Generative AI applications are built using these LLMs.

Machine Learning (ML): is a method for learning the rules of an algorithm based on existing data.

Machine learning model: is an algorithm that is built by learning patterns in existing data. For example, a machine learning model to predict house prices is constructed by learning from historical data on home prices. The model may learn that price increases with square footage, changes by neighborhood, and depends on the year of construction.

Microsoft Copilot: an AI-powered assistant integrated across Microsoft 365 applications that helps users draft content, summarize emails, analyze data, and automate tasks.

Model validation: methods to determine whether the outputs generated by a machine learning model are unbiased and accurate.

Public or consumer Generative AI Tool: free or third-party Generative AI tool not managed by the City.

Snowflake Cortex: a suite of built-in Generative AI and machine learning capabilities within the Snowflake data platform. Cortex allows users to securely analyze, summarize, and generate insights from data using large language models, all within the City's existing Snowflake environment.

Training data: The dataset that is used by a machine learning model to learn the rules.

GUIDELINE

Generative Artificial Intelligence Policy

POL-209

Purpose

The purpose of this policy is to set forth requirements City departments will observe when acquiring and using software that meets the definition of "generative artificial intelligence."

Scope

All City departments. Vendors, contractors, and volunteers who operate on behalf of the City are also subject to this policy.

Definitions

Generative Artificial Intelligence (Generative AI) is a class of computer software and systems, or functionality within systems, that use large language models, algorithms, deep-learning, and machine learning models, and are capable of generating new content, including but not limited to text, images, video, and audio, based on patterns and structures of input data. These also include systems capable of ingesting input and translating that input into another form, such as text-to-code systems.

While this policy document includes principles that apply to AI technologies generally, the policy statements apply only to generative AI systems.

Artificial Intelligence (AI) Principles

Principles describe general codes of conduct that represent the City's values and are aligned with our responsibilities to the residents we serve. These principles serve to guide City employees in their use of both generative and traditional AI technology. City employees shall adhere to the principles and requirements outlined in this policy, and will be held accountable for compliance with these commitments.

1. **Innovation and Sustainability:** The City values public service innovation to meet our residents' needs. We commit to responsibly explore and evaluate AI technologies, which will improve our services and advance beneficial outcomes for both people and the environment.
2. **Transparency and Accountability:** The City values transparency and accountability and understands the importance of these values in our use of AI systems. The City will ensure that the development, use, and deployment of AI systems are evaluated for and compliant with all laws and regulations applicable to the City prior to use, and will make documentation related to the use of AI systems available publicly.
3. **Validity and Reliability:** The City will work to ensure that AI systems perform reliably and consistently under the conditions of expected use, and that ongoing evaluation of system accuracy throughout the development and/or deployment lifecycle is managed, governed, and auditable, to the greatest extent possible.
4. **Bias and Harm Reduction and Fairness:** We acknowledge that AI systems have the potential to perpetuate inequity and bias resulting in unintended harms on Seattle residents. The City will evaluate AI systems through an equity lens, in alignment with our Race and Social Justice

commitments, for potential impacts such as discrimination and unintended harms arising from data, human, or algorithmic bias to the extent possible.

5. **Privacy Enhancing:** The City values data privacy and understands the importance of protecting personal data. We work to ensure that policies and standard operating procedures that reduce privacy risk are in place, and are applied to the AI system throughout development, testing, deployment, and use to the greatest extent possible.
6. **Explainability and Interpretability:** The City understands the importance of leveraging AI systems, models, and outputs that are easily interpreted and explained. We work to ensure all AI systems and their models are explainable to the extent possible, and that system outputs are interpretable and communicated in clear language, representative of the context for use and deployment.
7. **Security and Resiliency:** Securing our data, systems, and infrastructure is important to the City. We will ensure AI systems are evaluated for resilience and can maintain confidentiality, integrity, and availability of data and critical City systems, through protection mechanisms to minimize security risks to the greatest extent possible, in alignment with governing policy and identified best practices.

Policy

1. Acquisition of Generative AI Technology

1.1. Consistent with the City's standards for Acquisition of Technology Resources, City employees may be authorized to use pre-approved generative AI software tools or they may request a non-standard acquisition of generative AI software through Seattle IT's current request process.

1.2. Seattle IT shall review exception requests according to its current risk and impact methodology, which shall include specific review criteria for generative AI technology. Seattle IT shall either approve or deny a request according to its criteria.

1.3. The City's standard for technology acquisition applies to all technology, including free-to-use software or software-as-a-service tools.

1.4. If a technology that has already been approved for use in the City adds or incorporates generative AI capabilities, no additional approval is required to use those capabilities, however all other aspects in this policy apply to said use.

1.5. Seattle IT may revoke authorization for a technology that adds AI capabilities, or may restrict the use of those AI capabilities, if, in its judgment, those AI capabilities present risks that cannot be effectively mitigated to comply with this policy or other City policies.

2. Use of Generative AI Outputs

2.1. Outputs of Generative AI systems must be reviewed by humans prior to each use in an official City capacity ("Human in the Loop" or HITL). HITL review processes shall be documented by owning departments and shall demonstrate how the HITL review was conducted to adhere to the principles outlined in this document.

2.2. Documentation of HITL reviews shall be retained according to the appropriate records retention schedule.

3. Attribution, Accountability, and Transparency of Authorship

3.1. All **images and videos** created by Generative AI systems must be attributed to the appropriate Generative AI system. Wherever possible, attributions and citations to the City of Seattle should be embedded in the image or video (e.g., via digital watermark).

3.2. If **text** generated by an AI system is used substantively in a final product, attribution to the relevant AI system is required.

3.3. If a significant amount of **source code** generated by an AI system is used in a final software product, or if any amount is used for an important or critical function, attribution to the appropriate AI system is required via comments in the source code and in product documentation.

3.4. All attributions should include the name of the AI system used plus an HITL assertion (which should include the department or group who reviewed/edited the content).

> *Example: Some material in this brochure was generated using ChatGPT 4.0 and was reviewed for accuracy by a member of the Department of Human Services before publication.*

3.5. Departments shall interpret "substantive use" thresholds to be consistent with the principles outlined in this document as well as relevant intellectual property laws.

4. Reducing Bias and Harm

4.1. Generative AI systems may produce outputs based on stereotypes or use data that is historically biased against protected classes. City employees must leverage RSJI resources (e.g., the Racial Equity Toolkit) and/or work with their departmental RSJI Change Team to conduct and apply a Racial Equity Toolkit (RET) prior to the use of a Generative AI tool, especially uses that will analyze datasets or be used to inform decisions or policy. As per the objectives of the RSJ program, the RET should document the steps the department will take to evaluate AI-generated content to ensure that its output is accurate and free of discrimination and bias against protected classes.

5. Data Privacy

5.1. Use of generative AI tools shall be consistent with the principles and standards described in the City's Data Privacy Policy and Information Security Policy.

5.2. Unless suitable enterprise controls and data protection mitigations are in place, employees shall not submit data that is classified by the City's data classification guidelines as Confidential or Confidential with Special Handling, or that otherwise not considered to be acceptable to disclose to the public, shall not be submitted to Generative AI systems.

5.3. No City data or records, including inputs or prompts, are to be used for training or parameter-tuning for Generative AI models outside the City's control. AI technologies that cannot prevent City data or records from contributing to their language models may not be used by City employees.

6. Public Records & City Records Management

6.1. All records generated, used, or stored by Generative AI vendors or solutions may be considered public records and must be disclosed upon request.

6.2. All Generative AI solutions and/or vendors approved for City use shall be required to support retrieval and export of all prompts and outputs (either via exposed functionality or through vendor contract assurances).

6.3. City employees who use generative AI tools are required to maintain, or be able to retrieve upon request, records of inputs, prompts, and outputs in a manner consistent with the City's records management and public disclosure policies and practices.

Exceptions

Exceptions must be approved in advance through submission of a Seattle IT Exception Review Approval request in Service Hub. This can be submitted directly or with the assistance of Client Engagement personnel. Note: this section refers to exceptions to *this policy* as it relates to generative AI tools that are in use by the City. It does not refer to requests for acquisition of non-standard applications or technologies.

Non-compliance

The Chief Technology Officer (CTO) is responsible for compliance with this policy. Enforcement may be imposed in coordination with individual division directors and department leaders. Non-compliance may result in department leaders imposing disciplinary action, restriction of access, or more severe penalties up to and including termination of employment or vendor contract.

Related Standards and Policies

- City Privacy Policy [POL-202]
- Acquisition of Technology Resources [STA-209]
- Information Security Policy [POL-201]
- Data Classification Guideline [GUI-110]

Responsibilities

The policy will be maintained through the Data Privacy, Accountability and Compliance (DPAC) division, owned by the Director of DPAC and City of Seattle Chief Privacy Officer. Their responsibilities include creating and maintaining the generative AI risk and impact criteria and the documents and forms to support the exception review process for this technology.

Document Control

This policy shall be effective on 11/1/2023 and shall be reviewed annually.

Version	Content	Contributors	Approval Date
v 1.0	Initial Draft	Reviewer: Greg Smith – Chief Information Security Officer (CISO)	**10/23/2023**
	Final	**Approver:** **Jim Loter – Interim Chief Technology Officer (CTO)**	**10/23/2023**

TOWN OF LOS ALTOS HILLS
26379 Fremont Road
Los Altos Hills, CA 94022
Phone: (650) 941-7222
www.losaltoshills.ca.gov

Generative Artificial Intelligence Acceptable Use Policy

Approved by City Council Feb 15, 2024

1. Executive Summary

The Town of Los Altos Hills uses reliable, collaborative, and secure information technology solutions to support the efforts of staff in delivering high-quality services to its community.

As new, relevant local government technologies arise, the Town often assesses the value that each can deliver. In 2022, generative AI – creating meaningful output such as text and images from human prompts -- emerged as a powerful new tool in a wide range of contexts. The Town quickly recognized that this technology could be extraordinarily useful in the public sector while also acknowledging that it could create unacceptable risks.

The Town of Los Altos Hills will partner and collaborate with other government agencies in the region and beyond to maximize the benefits of generative AI and to reduce its risks such as security, privacy, and content issues.

This policy provides rules and guidance for Town staff to support responsible use of generative AI tools. The Town is committed to encouraging the use of emerging technologies that promote progress and innovation, increase organizational performance and quality service delivery, and serve the public good.

2. Version Control

Date	Version	What	Who
11/18/23	1.0	Draft policy framework	External consultant
11/24/23	2.0	First full draft of policy	External consultant
11/27/23	2.1	Review and minor modifications	City Manager
12/17/23	3.0	Incorporates feedback from tech committee	LAH Tech Committee
2/15/2024	3.0	Adopted by City Council	LAH City Council

3. Definitions

What	Description
User	Staff, contractors, or others using Generative AI for Town work purposes.
Town	The local government of The Town of Los Altos Hills, California.
Artificial Intelligence (AI)	A system that performs certain tasks that simulate those of a human.
Generative AI	A system that creates content such as text, audio, or images in response to human or computer inputs.

4. Purpose

The purpose of this policy is to set forth requirements and guidance that Town users must adhere to when acquiring and using solutions that meet the definition of generative artificial intelligence (AI).

5. Scope

While AI is a broad and deep topic, this policy is currently concerned with solutions that use a specific type of AI called generative AI. Examples include ChatGPT and Google Bard, as well as products from Stability.ai, Jasper.ai, and similar. This policy also acknowledges that the boundaries of such solutions can be complex to define, and the nature of the technology is changing at a high rate. In the event of uncertainty in how any part of this policy applies, staff should first seek guidance from a supervisor.

6. Benefits and Risks of Using Generative AI in Local Government

Benefits (not an exhaustive list):

a) Ability to analyze large amounts of information and then identify patterns, answer questions, make recommendations, and summarize findings.

b) Reduce time spent on regular tasks including the creation of documents such as memos, emails, job descriptions, and reports, and assist with rapid knowledge acquisition.

c) Improved decision-making by discovering and exploring scenarios, evaluating options, and analyzing relevant data.

d) Assist in the development of new content such as policies and procedures, website content, and social media posts, including rewording and improving grammar.

e) Provide community-facing chatbots that provide 24-hour access to information and services.

f) Translates text into different languages.

Risks (not an exhaustive list):

a) May create output that uses, for example, inappropriate, generic, and irrelevant language, bias, and tone. Translations too, may not be of an acceptable level of quality.

b) Information produced may not be accurate, which includes dates, places, names, events, and more.

c) May reveal personally identifiable or sensitive information, disclose confidential information, or release draft information to the public.

d) Reliance may reduce the important use of human empathy, personal contact, discretion, and judgement.

e) All input and output content may be subject to relevant public record laws.

f) Content produced may be subject to copyright laws and use could result in legal challenges.

7. Acceptable Uses of Generative AI

The range of uses of generative AI are wide and new capabilities are being frequently introduced. This list provides a small set of staff usage examples that can be used as guidance.

a) Creating an outline for written content. The generative AI content is used as a starting point, but the final product is created by staff. Examples include:

- Letters
- Emails
- Project documentation
- Speaker notes
- Presentation slides
- Social media posts
- RFPs and similar
- Web content
- Reports
- Procedures
- Policies
- Job descriptions
- Press releases

b) Copying a document into a generative AI tool for the purposes of summarizing and/or querying it.

c) Suggesting writing improvements.

d) Learning about and exploring topics.

e) Analyzing different types of data.

f) Idea generation.

g) Research.

8. Requirements for Using Generative AI

a) All generative AI tools used for work purposes require an account that is explicitly for Town use. Personal accounts are not permitted for work use.

b) Assume that all use of generative AI tools and content are subject to relevant public records requests and must adhere to existing Town data retention requirements where appropriate.

c) Only provide input content that is okay for public disclosure.

d) Staff are responsible for all outcomes of generative AI used for work purposes. Specifically, the consequences of generative AI use are a human responsibility and issues cannot be deferred to the software and system. Staff assume all responsibility for content produced by generative AI.

e) Conduct a fact check on all content that is produced by generative AI. Staff must validate all output and ensure that it is factually correct. Staff are responsible for managing the risks of AI hallucinations, i.e., the creation by AI of inaccurate or nonsensical content.

f) Provide attribution whenever generated content is published including images.

g) Log usage of generative AI for possible reference purposes later.

h) Do not use generative AI to generate code or audio at this time.

9. **Procurement of Generative AI Solutions**

a) To procure generative AI solutions, staff must follow the existing policies and procedures for acquiring hardware and software for the Town.

b) Staff acknowledge that the procurement process may be slower than normal as a risk and impact assessment may be required.

c) The policy and procedures for acquiring generative AI solutions also applies to free-to-use, freemium, open source, software-as-a-service (SaaS), and any other solution formats.

d) Approval to procure a generative AI solution is not required if the identical solution has already been formally approved.

e) Use of any generative AI solution may be restricted or revoked at any time if, in the judgement of the Town, the risk of use results in non-compliance in part or full, with this policy or any other Town policy, or other relevant policy, rule, or law.

10. Enforcement

Enforcement may be imposed in coordination with individual division directors and department leaders. Non-compliance may result in department leaders imposing disciplinary action, restriction of access, or more severe penalties up to and including termination of employment or vendor contract.

11. Training

The Town will occasionally provide training in generative AI and specific tools. For specific requirements, staff should work with their supervisor to request training following their normal learning needs process.

Khan Academy provides high-quality, free online training. Access via khanacademy.org and search for "generative AI"

For developing prompt engineering skills—the ability to provide good input for great generative AI output—the following site provides free, high-quality training: learnprompting.org

12. Contact and Oversight Information

This policy will be maintained through the City Manager's department, owned by the City Manager or designee. Their responsibilities include creating and maintaining the generative AI risk and impact criteria, making recommendations for policy updates, communicating updates on the Towns position on generative AI, and as a point of contact for escalating challenges and issues with the technology.

13. Document Control

This policy shall be effective on Feb 15, 2024 and shall be reviewed annually. A review and update may also be initiated in the event of substantive changes to generative AI and its associated solutions.

I-82.4 Artificial Intelligence Appropriate Use Policy
To: Department Heads
Subject: I-82.4 Artificial Intelligence Appropriate Use Policy

PURPOSE

This policy provides guidelines for City of Santa Cruz (City) employees and officials on the use of artificial intelligence (AI). The City is dedicated to the responsible and ethical use of AI technologies, aiming to enhance processes, improve services for residents, and empower employees to excel in their work. In response to the rapid development of AI tools, the City will consistently assess and revise this policy to ensure alignment with ethical, legal standards, and the latest advancements in AI as required.

This policy establishes a comprehensive, yet flexible, governance structure for AI systems used by, or on behalf of, the City. This policy enables the City to use AI systems for the benefit of the community while safeguarding against potential harms.

The key objectives of the AI Policy are to:

a. Provide guidance that is clear, easy to follow, and supports decision-making for the staff (full-time, part-time), interns, consultants, contractors, partners, and volunteers who may be purchasing, configuring, developing, operating, or maintaining the City's systems or leveraging AI systems to provide services to the City.
b. Ensure that when using AI systems, the City or those operating on its behalf, adhere to the Guiding Principles that represent values with regards to how AI systems are purchased, configured, developed, operated, or maintained.
c. Define roles and responsibilities related to the City's usage of AI systems.
d. Establish and maintain processes to assess and manage risks presented by AI systems used by the City.
e. Align the governance of AI systems with existing data governance, security, and privacy measures in accordance with Administrative Procedure Order ***(APO) I-82.1 Use of City of Santa Cruz's Electronic Systems and Tools.***
f. Define prohibited uses of AI systems.

SCOPE

This policy applies to any employee or official using or accessing the City's Technology Resources and Systems to access AI technologies deployed by the City or are involved in using AI tools or platforms on behalf of the organization. Employee as used in this policy is defined as an individual performing business activities under direct supervision of another City employee and includes full-time, part-time, and temporary employees, contractors, unpaid interns, and volunteers. Officials as used in this policy is defined as anyone in a position of official authority that holds a legislative or administrative position of any kind, whether appointed or elected.

DEFINITIONS

Artificial Intelligence (AI):
The City defines "artificial intelligence" or "AI" to be a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems use machine and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in

an automated manner; and use model inference to formulate options for information or action. The City defines an "AI system or technologies" to be any data system, software, hardware, application, tool, or utility that operates in whole or in part using AI.

Generative AI Tools
A category of AI tools designed to generate new content, data, or outputs based on patterns and information learned from existing examples. Generative AI tools are computer programs capable of many activities, including but not limited to completing general administrative office tasks, data analysis, programming, and image creation. While these tools can improve productivity, it is crucial to use them responsibly to comply with various laws, maintain data privacy and security, and uphold City values.

AI-Powered Search
A category of advanced search technology that integrates artificial intelligence (AI) to enhance traditional search engine functionalities, such as those found in platforms like Google and Bing. AI-powered search leverages techniques like machine learning, natural language processing (NLP), and deep learning.

POLICY

Guiding Principles
Staff should be open to responsibly incorporating AI tools into their work where it can be beneficial for making services better, more just, and more efficient. Each employee is responsible for using AI tools in a manner that ensures the security of sensitive information and aligns with City policies. Specific guiding principles include:

a. **Data Privacy and Security:** Comply with all data privacy and security standards such as Health Insurance Portability and Accountability Act (HIPAA), Criminal Justice Information Systems (CJIS), Internal Revenue Service (IRS), and the California Consumer Privacy Act (CCPA) to protect Personally Identifiable Information (PII), Protected Health Information (PHI), or any sensitive data in AI prompts.
b. **Treat AI prompts as if they were publicly visible online to anyone.** Treat AI prompts, data inputs, and outputs as if they are subject to the Freedom of Information Act and Public Records Act. Privacy is preserved in all AI systems by safeguarding personally identifiable information (PII) and sensitive data from unauthorized access, disclosure, and manipulation.
c. **Informed Consent:** Members of the public should be informed when they are interacting with an AI tool and have an "opt out" alternative to using AI tools available.
d. **Responsible Use:** AI tools and systems should be used ethically and always include human oversight to ensure accountability and proper decision-making.
e. **Continuous Learning:** Employees using AI tools are encouraged to educate themselves on effective and appropriate AI usage for their work.
f. **Avoiding Bias:** AI tools can create biased outputs. When using AI tools, develop AI usage practices that minimize bias and regularly review outputs to ensure fairness and accuracy, as you do for all content**.**
g. **Equity:** AI systems should support equitable outcomes for everyone. Bias in AI systems is effectively managed with the intention of reducing harm for anyone impacted by its use.
h. **Decision Making:** Do not use AI tools to make impactful decisions. Be conscientious about how AI tools are used to inform decision making processes.

i. **Accuracy:** AI tools can generate inaccurate and false information. Take time to review and verify AI-generated content to ensure quality, accuracy, and compliance with City guidelines and policies.
j. **Transparency:** The use of AI systems should be explainable to those who use and are affected by their use. The purpose and use of AI systems is proactively communicated and disclosed to the public to ensure transparency. When using this rapidly evolving technology, indicate when AI tools contributed substantially to the development of a work product. Use good judgement when considering citing use of AI tools. For example, when working with regulatory, legal, or life critical subjects, citations will be appropriate. When attributing AI usage in a work product, indicate the product and version used. Some example citations include:
    i. **Whole Document Example:** [Artificial Intelligence contributed to the development of this document using Google Bard, 2023, in compliance with the City of Santa Cruz Appropriate Use Policy.]
    ii. **Specific Citation Example**: OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]
    iii. **In text Example for specific information:** (OpenAI, 2023)
k. **Accountability:** Employees are solely responsible for ensuring the quality, accuracy, and regulatory compliance of all AI generated content utilized in the scope of employment. The City will regularly monitor and evaluate approved AI products to ensure they meet security and risk management criteria. Roles and responsibilities govern the deployment and maintenance of AI systems, and human oversight ensures adherence to relevant laws and regulations.
l. **Security & Safety:** AI systems maintain confidentiality, integrity, and availability through safeguards that prevent unauthorized access and use. Implementation of AI systems is reliable and safe, and minimizes risks to individuals, society, and the environment.

Prohibited Use of AI systems

a. **Unauthorized Access**: Do not use AI tools to access data or systems you are not authorized to use. Examples of systems that might be deemed unauthorized are AI systems that:
    i. utilize facial recognition,
    ii. make automated decisions without human oversight,
    iii. attempt to make emotional recognition,
    iv. provide social services decision-making,
    v. create hyper-realistic images, videos or audio that are indistinguishable from reality, or
    vi. perform automated image alteration without disclosure.
b. **Unlawful Activities:** AI tools must not be used for any illegal, harmful, or malicious activities. This includes activities that perpetuate unlawful bias, automate unlawful discrimination, and produce other harmful outcomes.
c. **Employment Decisions:** Do not use AI to make consequential decisions regarding hiring, benefits, discipline, or other sensitive matters in which bias could play a role, without additional verification and human oversight. Information gathered from AI, including but not limited to insight on an employee's work performance and behavior, shall not be used in any disciplinary decisions without additional verification and human oversight.

d. **Sensitive Data:** Do not enter internal, sensitive, or restricted data into any AI tool or service (e.g. PII, PHI, HIPAA, IRS, CJIS). No AI tool currently meets the City's security, privacy, and compliance standards for handling anything besides public data**.** However, in the event internal, sensitive, or restricted data may be utilized by an AI system, notice will be provided to all affected individuals within a reasonable time and manner.
e. **Automated decisions:** Do not use AI Tools to make fully automated decisions that do not require any meaningful human oversight but substantially impact individuals. AI should not be utilized as a sole means to prepare individualized personnel documents, including but not limited to, disciplinary documents and performance evaluations.
f. **Social scoring**: Do not use AI tools to track and classify individuals based on their behaviors, socioeconomic status, or personal characteristics.
g. **Behavioral manipulation:** AI tools must not be used as a cognitive manipulation tool of people or specific vulnerable groups.

Roles and Responsibilities

The Information Technology (IT) Director shall:

a. **Designate an IT Manager** to actively monitor and ensure AI systems are used in compliance with this policy as well as ***I-82.1 Use of City of Santa Cruz's Electronic Systems and Tools***. This will ensure AI systems are aligned with organizational policies and standards.
b. **Authorize an IT Manager** to conduct regular inspections of AI system usage within departments. If any department or its partners are found to be using AI systems inappropriately or outside policy guidelines, the IT Manager will have the authority to direct departments to modify or discontinue the use of such systems.
c. **Communicate policy updates** to all City departments whenever changes are made to this policy or the overarching City of Santa Cruz Technology Use Policies. This will help maintain consistency in policy adherence across all departments and ensure timely implementation of new directives.

The IT Manager shall:

a. **Supervise the enterprise security infrastructure**, ensuring that the proper mechanisms for monitoring and compliance with AI-related policies are in place. This includes continuous review of security controls and policy enforcement measures.
b. **Collaborate with City departments** to conduct an AI Risk Review. This review will evaluate the risks associated with the introduction, integration, and usage of AI systems. The objective is to identify potential vulnerabilities, data privacy issues, and ethical risks before adopting any AI system.
c. **Provide guidance on AI-related security practices** and offer training, where needed, to ensure departments understand the cybersecurity risks associated with AI tools and their proper usage.

The IT Department shall:

a. **Maintain and publish a list of approved AI tools** on the IT intranet site. This list will be reviewed and updated regularly to ensure departments have access to the latest approved tools.
b. **Facilitate AI tool requests** from City departments. If a department wishes to use an AI tool not on the approved list, they must contact the IT Department via helpdesk@santacruzca.gov. The IT Department will review the tool for security, compliance, and functionality before approving its use.

c. **Update and maintain guidelines for AI tool usage** and ensure that departments have clear documentation and processes in place to comply with the city's AI governance framework.
d. **Training:** Develop and provide training resources to the Human Resources department to be included in the new employee orientation training.

City Departments shall:

a. **Adhere to this AI policy** and any future updates, as well as ***I-82.1 Use of City of Santa Cruz's Electronic Systems and Tools***. Departments must annually check for compliance with these documents to ensure continuous alignment with updated policies and technological changes.
b. **Seek approval from the IT Department** prior to implementing any unapproved AI system. This ensures that all AI systems are vetted for security, ethical implications, and compliance with city policies before they are deployed.
c. **Use an official City of Santa Cruz email address** when registering or creating accounts with AI service providers to maintain security and accountability. This will also facilitate easier monitoring and traceability in case of any incidents related to AI usage.
d. **Report AI-related concerns or potential issues** immediately to the IT Department. This includes any suspicious activity, security breaches, or ethical concerns stemming from the use of AI tools, ensuring timely mitigation and resolution.

NO RIGHT OF PRIVACY

See I-82.1 Use of City of Santa Cruz's Technology Resources and Tools.

VIOLATIONS OF POLICY

See I-82.1 Use of City of Santa Cruz's Technology Resources

City of Portland Core Values:
Anti-racism | Equity | Transparency | Communication | Collaboration | Fiscal Responsibility

BTS – 4.04 Artificial Intelligence Use and Governance

Purpose

This administrative rule establishes requirements for the acquisition, development, and deployment of artificial intelligence (AI) systems and services by City bureaus and service areas. These requirements ensure the responsible, ethical, consistent, and secure use of AI technologies in alignment with:

- The City of Portland's Core Values; and
- The City's existing administrative rules regarding information security and acceptable use of technology.

The key objectives of this Rule are to:

- Provide clear, practical guidance which supports decision-making for City-authorized users who may be purchasing, configuring, developing, operating, using, or maintaining the City's AI systems or who use AI systems to provide services to the City.
- Define roles and responsibilities related to the oversight and responsible use of City AI systems.
- Inform all City-authorized users that, when using AI-enabled systems, they must adhere to City's AI Usage Guidelines for appropriate use of AI systems.
- Align the oversight of AI system security with existing Information Security and privacy requirements in the related BTS Information Security Administrative Rule.
- Align the acceptable use of AI systems with existing related HR administrative rules as referenced in Guidance for this rule.
- Establish and maintain processes to assess and manage risks and benefits presented by the implementations of AI systems in the City.
- Promote transparency by communicating the purpose and use of AI systems to staff and the public, including through inventories or summaries of approved uses, consistent with City privacy and security requirements.
- Outline prohibited uses of AI systems.

- Ensure compliance with public records inspection and retention laws

The City Code authorizes the Bureau of Technology Services (BTS) Director position held by the Chief Information Officer to manage and establish this citywide policy and standards. BTS is responsible for ensuring compliance with this Administrative Rule and for making decisions regarding risk assessments, appropriate use, vendor requirements, and operational oversight of AI products and services within the City.

Key Definitions and Terminology

Artificial Intelligence (AI): Machine-based systems that, based on human-defined objectives or parameters, can make predictions, generate recommendations or decisions, retrieve, synthesize, and analyze data, and perform other tasks that influence real or virtual environments. AI systems use human and machine generated inputs to interpret environments, transform those interpretations into models, and produce outputs (such as options for information or action) through automated inference.

Algorithm: A defined set of instructions or logical steps used by a computer program (including AI agents) to transform inputs into outputs.

AI systems: Any software, hardware, service, or process that uses algorithms and data to automatically generate outputs, including predictions, recommendations, or decisions, that augment or replace human judgment. This includes systems used to automate processes, analyze large data sets, or otherwise perform tasks traditionally requiring human intelligence.

City-authorized users: People including full-time and part-time employees, contractors, vendors, volunteers, interns, consultants, partners, and other authorized parties.

Generative AI (GenAI): A sub-class of AI systems that use machine learning models (including large language models, deep learning, and related algorithms) to generate new content. This includes, but is not limited to, text, images, video, audio, code, and other media produced in response to input data or "prompts." Generative AI may also transform inputs into different formats, such as converting text into computer code or creating graphics, images, or video from text prompts.

Eligibility or Applicability

This Rule applies to:

All AI systems and services deployed by the City or operated on the City's behalf that process City data, directly support City operations, or interact with City staff or the public. It does not extend to incidental or internal vendor or contractor uses of AI unrelated to City work (e.g., a vendor's internal HR system, spam filtering, or productivity tools). This exclusion for incidental or internal vendor or contractor use does not apply when a vendor's AI system processes, stores, or analyzes City data for any purpose beyond routine IT operations such as system maintenance, security monitoring, or spam filtering.

All City-authorized users (including full-time and part-time employees, contractors, vendors, volunteers, interns, consultants, partners, and other authorized parties) who may be purchasing, configuring, developing, operating, using, or maintaining the City's AI systems, or who use AI systems as part of the services they deliver to the City.

Vendors who access or process City information while providing goods or services. Vendor use of City data with AI technologies is limited to the purposes expressly authorized in their contract with the City and must comply with all applicable contractual, cybersecurity, and data protection requirements. City information may not be used for training, testing, or improving vendor AI models unless explicitly permitted in writing by the City.

Roles, Responsibility, Process and Procedures

Roles and Responsibilities

The **Chief Information Officer (CIO)** is responsible for managing City technology resources, policies, projects, and services. These responsibilities include AI systems and services. The CIO coordinates these across City bureaus and service areas. The CIO and BTS will own and maintain this policy and notify City bureaus and service areas when updates are released.

The **Chief Information Security Officer (CISO)** ensures AI systems are used in accordance with BTS Administrative Rule and all applicable compliance regulations. The CISO oversees enterprise cybersecurity infrastructure and cybersecurity operations, and updates security policies, procedures, standards, guidelines, and incident response measures as needed.

The **Chief Procurement Officer (CPO**) oversees the procurement of AI systems in alignment with existing IT procurement and governance processes, administrative rules, and public procurement laws. The CPO, in collaboration with the City Attorney, is responsible for requiring

vendors to comply with applicable City policies and standards through contractual agreements.

The **City Administrator, or designee,** may, at their discretion, inspect the usage of AI systems and require a bureau or service area to alter or cease its usage of AI systems, or the usage of AI systems by a partner on behalf of the City.

The **Bureau of Technology Services (BTS)** will evaluate AI products and their alignment with City IT strategy, risk assessments, and procurement procedures. BTS will provide ongoing oversight of AI services and products in the City.

Identifying Artificial Intelligence

City employees and other authorized users are required to consider whether a proposed product or service includes AI functionality when making or evaluating technology use or purchase proposals. To assist in this evaluation, the following questions serve as an initial screening:

- Does the technology use data to provide predictions, recommendations, insights, or decisions?
- Does the technology generate language, graphics, data, or computer code based on human inputs or "prompts"?
- Does the technology augment or replace human decision-making?
- Does the vendor offering the technology use words such as "personalized," "tailored," or "adaptive" in its marketing?

A "yes" to any of these questions may indicate that the technology is an AI system and may therefore fall within the scope of this rule.

When uncertain, staff should consult BTS, which will make the final determination. If there are questions or disagreements about whether a product or service qualifies as an AI system, BTS will make the final determination.

Procurement, Configuration, Operation and Use

Procurement of AI Products and Services

The requestor, typically a bureau or service area IT Manager or equivalent role, must submit a business case to BTS for an initial risk assessment prior to initiating procurement of an AI system. The risk assessment will evaluate the potential impacts and risks of the proposed implementation.

1. The CPO and CISO are responsible for developing and maintaining the City's AI System Procurement Requirements and for ensuring that contractors comply with these requirements.
2. The CPO will require vendors to complete the City's hosted service questionnaire before finalizing an AI-related procurement. The questionnaire will be updated to include AI-specific requirements and disclosures, including whether City data will be used for model training, fine-tuning, evaluation or improvement. Vendors must provide technical documentation describing data handling, model behavior, and any adaptive or learning components of the system. The CPO will collect and maintain this documentation as part of the procurement and contracting record. For AI systems developed by a bureau or project team, equivalent disclosures and documentation must be provided.
3. Vendor use of City data for training, testing, or improving AI models is prohibited unless explicitly authorized in writing by the City and governed by applicable contractual, cybersecurity, and data protection requirements. Contracts must include audit or verification rights, proportional to risk, that allow the City to confirm vendor compliance with agreed-upon data use, security, and AI-specific requirements.
4. When bureaus become aware of AI being added to an existing City system, the responsible bureau must notify BTS. BTS, in consultation with CPO and the CISO as needed, will review the new features and may require updated questionnaires, disclosures, documentation or other actions. BTS will apply the same risk model used for new AI systems and may require mitigations, impose conditions, or prohibit use if risks cannot be reduced to acceptable levels.

Configuration, Operation, and Use

City-authorized users must abide by this Administrative Rule in the implementation and use of AI systems.

1. City-authorized users must comply with the City's Human Resources Administrative Rule for Information Technologies.
2. City-authorized users must follow the City's AI Usage Guidelines when using AI systems to generate content for City or public use, including requirements for review, validation and disclosure of AI involvement.

3. All implementations and use of AI systems must comply with City Information Security rules as specified in BTS Information Security Administrative Rule.
4. All content and services that use AI systems to generate their outputs must provide language access consistent with the City's Language Access Policy referenced in Guidance for this Rule and compliance with Title VI of the Civil Rights Act of 1964.
5. AI systems that interact with the public or influence public-facing services must clearly communicate when AI is used, consistent with the transparency expectations outlined in the City's AI Usage Guidelines and applicable risk-based review requirements.
6. Use, configuration, and operational safeguards must reflect the system's risk level under the City's AI Risk Model, including any additional oversight, controls, or disclosures required for higher-risk uses.

Prohibited Uses

AI systems, tools, and services provided or authorized by the City of Portland must not be used for any criminal, malicious, discriminatory, or otherwise unauthorized purpose. Exceptions must be specifically approved by the City Administrator, or designee, and no exception may authorize conduct that is unlawful, unethical, or contrary to City policy.

Examples of prohibited uses include, but are not limited to:

- Illegal or malicious activities, including fraud, cyberattacks, or other unlawful conduct.
- Circumventing security, privacy, or intellectual property protections.
- Harassment, surveillance, or profiling outside of lawful and approved purposes (e.g., real-time or covert biometric identification in public spaces, such as facial or gait recognition).
- Automated decision-making without an appropriate level of human review, proportionate to the risk, when outcomes could significantly affect an individual's health, safety, rights, or financial well-being (e.g., decisions about an individual's employment, or specific law enforcement actions). Automated tools may assist with efficiency, but final determinations in these areas require human review.
- Manipulating or exploiting individuals or groups, including through subliminal, coercive, or deceptive techniques (e.g., emotion recognition or behavior-shaping systems).

- Assigning social scores or rankings to individuals based on personal characteristics, socioeconomic status, or behaviors. This prohibition does not apply to the City's use of aggregated or de-identified data (such as census or household-level data) for legitimate statistical analysis, planning, or service delivery.
- Producing or distributing harmful, misleading, or deceptive content in the name of the City or while using City resources.

Principles and Risk

Principles of AI Use

These principles describe the City's values regarding how AI systems are purchased, configured, developed, operated, used and maintained and are further defined under the Guidance for this Rule.

- Human-Centered Design and Use,
- Security and Safety,
- Privacy,
- Transparency,
- Equity,
- Accountability,
- Effectiveness, and
- Workforce Empowerment,

Risk Management

AI implementations pose unique risks in addition to those inherent in all information technology. Therefore, it is important to have a risk assessment and management model that evaluates risks based on the specific use case.

BTS will use a risk-based approach to conduct an initial AI risk assessment. This AI Risk Model assessment is referenced in the Guidance for this Rule and is further detailed in the City's AI Usage Guidelines. It will determine the extent of controls, compliance criteria, and oversight required for a particular AI product, service, or use case. BTS will support this assessment by coordinating privacy, equity, and surveillance reviews as applicable and by applying the governance processes established under the City's AI governance program.

The initial AI risk assessment does not replace or take precedence over other required risk assessments for new technologies. These include, but are not limited to, information security, privacy, financial, and legal risk assessments. Such assessments will be performed by their respective

	responsible parties in partnership with BTS and will be integrated into the City's broader AI governance workflow.
Resources	Refer to BTS 4.04 **Guidance** (*attached as separate document and to be linked once rule is approved and published*) for additional details, references to related Administrative Rules, contact information and other resources including the City's AI Usage Guidelines.
Administrative Rule Record	Adopted by City Administrator as a Final Rule, effective March 6, 2026

Artificial Intelligence (AI) Policy 1.7.12

PURPOSE

This policy establishes a governance structure that allows the City of San José (hereafter referred to as "City") to utilize Artificial Intelligence (AI) and AI systems (systems) while providing the necessary safeguards for purposeful and responsible use.

The key objectives of the AI Policy are to:

- Provide guidance that is clear, easy to follow, and supports decision-making for the staff, interns, consultants, contractors, partners, and volunteers who may be purchasing, configuring, developing, using, maintaining, or leveraging AI to provide services to the City;
- Ensure that the use of AI systems adheres to the Guiding Principles with regard to how systems are purchased, configured, developed, operated, or maintained;
- Define roles and responsibilities related to the usage of AI;
- Establish and maintain processes to assess and manage risks presented by AI;
- Align the governance of AI with existing data governance, security, and privacy measures in accordance with the City's Information and Systems Security Policy and City Council's Digital Privacy Policy;
- Define prohibited uses of AI systems;
- Establish "sunset" procedures to safely retire systems that no longer meet the needs of the City; and
- Define how AI may be used for legitimate purposes in accordance with applicable local, state, and federal laws, and existing agency policies.

AI systems and the data contained therein will be purchased, configured, developed, operated, and maintained using the City's AI Handbook, which will be managed by the Chief Information Officer (CIO).

SCOPE

This policy applies to:

1. All systems deployed by the City; and
2. Staff, interns, consultants, contractors, partners, and volunteers who may be purchasing, configuring, developing, using, or maintaining the AI or who may be leveraging systems to provide services to the City (collectively referred to as "users").

GUIDING PRINCIPLES FOR RESPONSIBLE AI SYSTEMS

These principles describe the City's values with regard to how AI systems are purchased, configured, developed, used, or maintained.

1. **Effectiveness:** Systems are reliable, meet their objectives, and deliver precise and dependable outcomes for the utility and contexts in which they are deployed;
2. **Transparency:** The purpose and use of systems is proactively communicated and disclosed to the public. A system, its data sources, operational model, and policies that govern its use are understandable and documented;
3. **Equity:** Systems deliberately support equitable outcomes for everyone. Bias in systems is

effectively managed with the intention of reducing harm for anyone impacted by the system's use;

4. **Accountability:** Roles and responsibilities govern the deployment and maintenance of systems, and human oversight ensures adherence to relevant laws and regulations;
5. **Human-Centered Design:** Systems are developed and deployed with a human-centered approach that evaluates AI powered services for their impact on the public;
6. **Privacy:** Privacy is preserved in all AI systems by safeguarding personally identifiable information (PII) and sensitive data from unauthorized access, disclosure, and manipulation in accordance with the City Council's Digital Privacy Policy;
7. **Security & Safety:** Systems maintain confidentiality, integrity, and availability through safeguards in accordance with the City's Information and Systems Security Policy. The integrity of information into and out of the City is maintained in light of fake AI-generated content. Implementation of systems is reliable and safe, minimizing risks to individuals, society, and the environment; and
8. **Workforce Empowerment:** Staff are empowered to use AI in their roles through education, training, and collaborations that promote participation and opportunity.

RESPONSIBILITIES

Several roles are responsible for enforcing this Policy, outlined below.

- The Information Technology Department Director / Chief Information Officer (CIO) is responsible for directing technology resources, policies, projects, services, and coordinating the same with other departments. The CIO shall designate the City Information Security Officer (CISO) and City Digital Privacy Officer (CDPO) to actively ensure the security, resilience, privacy, and policy compliance of the systems used by the City.
- The CISO and CDPO are responsible for recommending updates to this policy and the AI Handbook.

POLICY

When purchasing, configuring, developing, using, or maintaining AI systems, users will:

1. Uphold the Guiding Principles for AI systems outlined above**;**
2. Conduct an AI Review to assess the potential risk of the AI system. The CDPO or designee is responsible for coordinating review of AI systems used by the City as detailed in the AI Handbook;
3. Obtain technical documentation about AI systems. The Finance Department, or other department overseeing the purchase of an AI system, is responsible for requiring vendors to disclose AI usage and to provide technical documentation (e.g., via the AI FactSheet as defined in the Terms and Definitions section, below) at the request of the CDPO; and
4. In the event of an incident involving the use of the AI system, follow the City's AI Incident Response Plan in accordance with the Information and Systems Security Policy. The CISO is responsible for overseeing the security practices of AI systems used by or on behalf the City.

Additionally, Finance is required to ask vendors to disclose the use of AI in procurement solicitations and to comply with the Requirements for AI Systems upon the request of the CDPO or designee.

Prohibited Uses

The use of certain AI systems is prohibited due to the sensitive nature of the information processed and severe potential risk. This includes, but is not limited to, the following prohibited purposes:

- Real-time and covert biometric identification;
- Emotion analysis, or the use of computer vision techniques to classify human facial and body movements into certain emotions or sentiment (e.g., positive, negative, neutral, happy, angry, nervous);
- Fully automated decisions that substantially impact the rights or safety of individuals with no meaningful human oversight;
- Social scoring, or the use of AI systems to track and classify individuals based on their behaviors, socioeconomic status, or personal characteristics; and
- Cognitive behavioral manipulation of people or specific vulnerable groups.

If staff become aware of an instance where an AI has caused harm, staff must report the instance to their supervisor, the CDPO, and the Office of Employee Relations no later than 24 hours after discovery.

Sunset Procedures

If an AI system operated by the City or on its behalf ceases to provide a positive outcome to the City as determined by the staff or CDPO, then the City must halt the use of that system unless express exception is provided by the CIO. If the abrupt cessation of the use of that AI system would significantly disrupt the delivery of services, a gradual phased out approach must be approved by the CIO before sunsetting. All measures to minimize the impact and recovery must be considered in the termination or phase out protocol, including but not limited to:

- Ownership and future access of data;
- Portability of the AI model, algorithm, and/or data; and
- Impact to services, users, and residents.

Public Records

The City must consider applicable public records laws before implementing an AI system and must comply with the City's Open Government and Ethics Provisions and the California Public Records Act. More information can be found in City's Administration Policy Manual 6.1.4 Open Government Policy.

Policy Enforcement

All employees and agents of the City, whether permanent or temporary, interns, volunteers, contractors, consultants, vendors, and other third parties operating AI systems on behalf of the City are required to abide by this Policy and the associated AI Handbook.

VIOLATIONS OF THE AI POLICY

Violations of any section of the AI Policy, including failure to comply with the AI Handbook, may be

subject to disciplinary action, up to and including termination. Violations made by a third party while operating an AI system on behalf of the City may result in a breach of contract and/or pursuit of damages. Infractions that violate local, state, federal or international law may be remanded to the proper authorities.

TERMS AND DEFINITIONS

Artificial Intelligence: "Artificial intelligence" or "AI" is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action.

Algorithm: A series of logical steps through which an agent (typically a computer or software program) turns particular inputs into particular outputs.

System: Any software, sensor, or process that uses AI to automatically generate outputs including, but not limited to, predictions, recommendations, or decisions that augment or replace human decision-making. This extends to software, hardware, algorithms, and data generated by these systems used to automate large-scale processes or analyze large data sets.

AI Fact Sheet: A template that captures the "nutrition facts," or essential technical details, of an AI system. Vendors are expected to complete the AI Fact Sheet during the procurement process. The AI Fact Sheet is a critical document that provides technical information needed to adequately understand, evaluate, and use AI systems. Maintained by the Information Technology Department.

Approved:

/s/	6/28/2024
Khaled Tawfik Information Technology Department Director / Chief Information Officer	Date

Approved for posting:

/s/	6/28/2024
Jennifer A. Maguire City Manager	Date

Artificial Intelligence (AI) Policy

Scope: CITYWIDE

Policy Contact:
Information Technology Department
(916) 808-7111
itservicedesk@cityofsacramento.org

Table of Contents:

Supersedes:
N/A - New

Reviewed/Effective:
01/09/2026

I. Purpose

This policy establishes a comprehensive, yet flexible, governance and management of Artificial Intelligence (AI) systems used by, or on behalf of, the City of Sacramento (City). This policy enables the City to use AI systems for the benefit of the City and community while safeguarding against potential harm.

The key objectives of the AI Policy are to:

A. Provide guidance that is clear, easy to follow, and supports decision-making for all end users who purchase, configure, develop, implement, maintain, operate, or use the City's AI systems.

B. Ensure that end users adhere to Section 4, Guiding Principles.

C. Define roles and responsibilities related to the usage of AI systems.

D. Establish and maintain processes to assess, manage, and mitigate risks presented by AI systems.

E. Align the governance of AI systems with existing data governance, security, and privacy measures in accordance with the City's Information Security Policy.

F. Define prohibited uses of AI systems.

G. Define how AI systems may be used for authorized City purposes in accordance with applicable local, state, and federal laws.

II. Scope

This policy applies to:

A. All systems with AI capabilities deployed by the City; and

B. End users.

III. Definitions

A. **Algorithm**: a series of logical steps through which an agent (typically a computer or software program) turns particular inputs into particular outputs.

B. **Artificial Intelligence (AI)**: a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems use machine-based and human-based inputs to:

1. Perceive real and virtual environments.

2. Abstract such perceptions into models through analysis in an automated

manner.

3. Use model inference to formulate options for information or action.

C. **Artificial Intelligence (AI) System**: Any system, software, sensor, or process that automatically generates outputs including, but not limited to, predictions, recommendations, or decisions that augment or replace human decision-making. This extends to software, hardware, algorithms, and data generated by these systems, used to automate large-scale processes or analyze large data sets.

D. **End Users**: City employees, administrators, contractors, consultants, vendors, volunteers, temporary agency employees, student interns, or any other person who uses or is provided access to Information Technology resources in accordance with this policy.

E. **Generative Artificial Intelligence (GenAI)**: A type of AI that is algorithmically trained on one or more large datasets and designed to generate new and unique data (e.g., text, pictures, video) in response to a prompt (generally questions, instructions, images, or video) input by the user.

F. **Human in the Loop**: The application of human oversight, intervention, or review throughout the various stages of an AI system's decision-making process. It provides for human oversight to ensure the AI's actions will align with desired outcomes and guiding principles for responsible use of AI.

IV. Guiding Principles for Responsible AI Systems

These principles describe the City's values regarding how AI systems are purchased, configured, developed, operated, implemented, or maintained.

A. **Human-Centered Design**: AI systems are developed and deployed with a human-centered approach that evaluates AI-powered services for their impact on the public.

B. **Security & Safety**: AI systems maintain confidentiality, integrity, and availability through safeguards that prevent unauthorized access and use. Implementation of AI systems is reliable and safe and minimizes risks to individuals, society, and the environment.

C. **Privacy**: Privacy is preserved in all AI systems by safeguarding personally identifiable information (PII) and sensitive data from unauthorized access, disclosure, and manipulation.

D. **Transparency**: The purpose and use of AI systems is proactively communicated and disclosed to the public. An AI system, its data sources, operational model, and policies that govern its use are understandable and documented.

E. **Equity**: AI systems are designed to ensure fair outcomes for everyone. Bias in AI systems is managed to minimize any potential harm to those affected by their use.

F. **Accountability**: Roles and responsibilities govern the deployment and maintenance of AI systems, and human oversight ensures adherence to relevant laws and regulations.

G. **Effectiveness**: AI systems are reliable, meet their objectives, and deliver precise and dependable outcomes for the utility and contexts in which they are deployed.

H. **Workforce Empowerment**: End users are empowered to use AI in their roles through education, training, and collaborations that promote participation and opportunity.

V. Roles & Responsibilities

Several roles are responsible for enforcing this policy, as outlined below.

A. **Chief Information Officer (CIO)** is responsible for:

1. Directing City technology resources, policies, projects, and services, and coordinating the same with other City departments.
2. Designating the Information Security Officer (ISO) to actively ensure AI systems are used in accordance with the Information Security Policy.
3. Designating the Artificial Intelligence Officer (AIO) to actively ensure AI systems are used in accordance with this policy and the Information Security Policy.
4. Updating this policy as needed.
5. Performing such other duties necessary to implement this policy.

B. **Information Security Officer (ISO)** is responsible for overseeing the enterprise cybersecurity infrastructure, cybersecurity operations, and updating information security policies, procedures, standards, and guidelines.

C. **The Artificial Intelligence Officer (AIO)** is responsible for:

1. Overseeing the enterprise digital privacy practices, data processing practices, and responsible usage of technology in compliance with this policy; and monitoring policy compliance.
2. Overseeing the privacy practices of AI systems used by or on behalf of City departments.
3. Coordinating the review of AI systems.

D. **Charter Officers and Department Directors** are responsible for:

1. Ensuring that their respective departments engage the AIO before seeking to procure new technology and data initiatives that involve an AI system.

2. Ensuring that end users within their respective departments receive and review this policy through the City's Learning Management System (LMS).

E. **The City Attorney's Office** is responsible for advising of any legal issues or risks associated with AI systems usage by or on behalf of City departments.

VI. Policy

A. **AI Consideration**

When purchasing, configuring, developing, operating, implementing, or maintaining AI systems, the City will:

1. Uphold Section 4, Guiding Principles for Responsible AI Systems.

2. Follow Procurement of Goods for items requiring IT Department review in procuring and implementing AI systems and solutions.

3. Conduct an AI Review to assess the potential risk(s) of AI systems.

4. Ensure all content generated by Generative AI is reviewed by a human to ensure accuracy and eliminate bias prior to being used, also known as "human in the loop."

5. Obtain technical documentation about AI systems.

6. In the event of a cybersecurity incident involving the use of the AI system, the City will follow an established incident response procedures. The ISO is responsible for overseeing the cybersecurity practices of AI systems used by or on behalf of the City.

7. Consistent with its obligations under the Meyers-Milias-Brown Act (MMBA), the City shall endeavor to provide notice to Recognized Employee Organizations (REOs) of the introduction of new AI technologies that the City reasonably anticipates may have more than a de minimis impact on employees' working conditions. Upon request, the City shall meet and confer regarding such impacts. Failure to provide prior notice shall not preclude the REO from requesting to meet and confer upon learning of such implementation. Nothing in this provision shall be construed to limit the City's authority under applicable MOU City Rights articles or other provisions of law.

B. **Use of AI and/or GenAI**

The use of AI and/or GenAI systems by end users shall be limited to official work-related purposes, and end users shall only access and use GenAI systems for which they have been authorized and received proper training.

Any function carried out by an end user using AI and/or GenAI is subject to the same laws, rules, and policies as if carried out without the use of AI and/or GenAI. The use of AI and/or GenAI does not permit any law, rule, or policy to be bypassed or ignored.

End users shall use AI-generated content as an informational tool and not as a substitute for human judgment or decision-making. End users shall not represent AI-generated content as their own original work.

AI-generated content is considered draft material only and shall be thoroughly reviewed prior to use. Before relying on AI-generated content, end users should:

1. Obtain independent sources for information provided by AI and/or GenAI and take reasonable steps to verify that the facts and sources provided by AI and/or GenAI are correct and reliable.

2. Review prompts and output for indications of bias and discrimination and take steps to mitigate its inclusion when reasonably practicable.

C. **Privacy Considerations for Public Safety Employees**

Information not otherwise available to the public, including data reasonably likely to compromise an investigation, reveal confidential law enforcement techniques, training, or procedures, or risk the safety of any individual if it were to become publicly accessible, confidential medical or patient information shall not be input into a GenAI system unless contractual safeguards are in place to prevent such information from becoming publicly accessible. End users should instead use generic unidentifiable inputs, such as “suspect,” “victim,” or “patient” and hypothetical scenarios whenever possible.

Protected information shall only be input into GenAI systems that have been approved for such use and comply with applicable privacy laws and standards.

VII. Prohibited Uses

A. AI Systems and IT resources may not be used to violate laws, policies, or contracts. End users may not use AI tools or services to generate content that enables harassment, threats, defamation, hostile environments, stalking, or discrimination.

B. If an employee becomes aware that an AI system is malfunctioning or producing inaccurate, misleading, discriminatory, or disruptive results (e.g., generating errors

in communication, calculations, or decision-making), they must promptly report the issue to their supervisor and the CIO.

C. Current employees and external job candidates are prohibited from utilizing AI to assist them with answering job interview questions. Utilizing AI for this purpose may result in the job candidate being disqualified from the recruitment.

D. End users shall not use AI and/or GenAI systems to rationalize a law enforcement decision, or as the sole basis of research, interpretation, or analysis of the law or facts related to a law enforcement contact or investigation.

E. End users are prohibited from inputting work-related data, including information learned solely in the scope of their employment with the City, into publicly available AI and GenAI systems. This restriction applies unless the system has been approved by the CIO or an authorized designee for the intended use. This policy is in place to ensure data security and compliance with City policies.

VIII. Violations of the AI Policy

A. City employees and volunteers who violate this policy may be subject to disciplinary action, up to and including termination of employment.

B. Contractors or vendors who violate this policy may be subject to appropriate actions as defined in contractual agreements and other legal remedies available to the City.

City of SACRAMENTO

Charter Officer Review and Acknowledgement

ARTIFICIAL INTELLIGENCE POLICY

(Signature by all Charter Officers is not a requirement for policy adoption)

Maraskeshia Smith

Maraskeshia Smith (Jan 14, 2026 19:26:31 PST)

City Manager 01/14/2026

Gus Martinez (Jan 9, 2026 11:28:37 PST)

City Attorney 01/09/2026

City Clerk 01/12/2026

City Treasurer 01/09/2026

City Auditor 01/20/2026

City of Anaheim

Administrative Regulation

CHAPTER 1 - ADMINISTRATIVE

Subject:	**Artificial Intelligence Appropriate Use Policy**	**A.R.**	**158**
		Issue Date:	**June 10, 2026**
		Revised:	
		Page	**(1) of (6)**

Purpose:
The purpose of this administrative regulation is to:

Establish clear guidelines for the responsible and ethical use (otherwise referred to as appropriate use) of artificial intelligence (AI) tools by city users. As new technologies such as AI become integrated into our public sector work, it is essential to ensure that their use aligns with appropriate legal standards, protects privacy, maintains public trust, and promotes the city's operational efficiency. This administrative regulation provides the framework for city users to use AI responsibly while ensuring compliance with these standards. AI shall only be used as a resource to support decision-making processes and tasks, not to replace human judgment or accountability, particularly for tasks, projects and assignments that involve critical decision making or outcomes.

Scope:
This administrative regulation applies to:

All city users of AI tools while conducting official city business, including employees and officials. It covers the use of AI tools, including those that generate text, images, videos, or other content based on data that is inputted, as well as AI models that are trained on or process public or private information.

Definitions:

Artificial Intelligence (AI): Any tool, software, platform, system, or application, whether provided by an external source, or created by the city, that can provide a response, produce content (e.g., text, images, code, etc.) or provide a service, based on information, instructions, or prompts provided by a human user.

City Data: Any data created, managed, or used by the city, including but not limited to documents, emails, reports, and databases containing personal, public, or confidential information.

Authorized Use of AI:
AI tools may be used for city-related tasks in the following circumstances, with users independently verifying factual statements, citations, legal authorities, calculations, and recommendations generated by AI systems before relying upon or distributing such information:

1. **Task Automation**: Streamlining repetitive tasks such as report generation, summarization, and customer service interactions.

2. **Research and Analysis**: Assisting in data processing, generating initial drafts of documents, research, and performing analysis.

3. **Creative Design**: Generating content for marketing, public relations, and digital communication, including social media posts and informational materials.

4. **Non-sensitive Data**: Use of public, non-sensitive data inputted into approved city AI tools to generate analysis or content is acceptable if it adheres to city policies herein and any relevant laws. Any sensitive data entered into an AI tool must be secure and not open to public viewing.

5. **City Approved Tool Use Only**: Only city approved AI tools may be used by city users. All non-approved tools must receive approval from the department, IT Manager, and from the city manager. Any new or experimental AI tools must be approved by the city before city users are authorized to use them. This includes review and approval from IT. The city is cognizant that public-facing browser-based AI tools, such as ChatGPT exist, and while permissible to utilize, users are not-permitted to enter sensitive, personal or confidential information.

6. **Examples**: Some examples of authorized uses include drafting city council agenda summaries, analyzing data, responding to service requests or creating memos summarizing research.

Prohibited Use of AI:
In addition to not using AI for any purposes that may violate any laws, the following uses of AI are prohibited:

1. **Generating Misinformation**: Using AI to produce or disseminate false, misleading, or biased information.

2. **Personal Use**: Using AI for personal use not related to city business.

3. **Infringing on Intellectual Property**: Using AI to produce work that plagiarizes or violates copyrights, trademarks, or other intellectual property rights.

4. **Processing Sensitive or Confidential Personal Information**: AI should not be used to process sensitive or confidential personal data unless explicit approval is obtained from the department head, and safeguards are in place. This includes only entering sensitive or confidential information into secure AI tools created for the city and not open to public viewing.

5. **Privileged or Confidential Information**: No user shall input privileged communications, attorney work product, litigation strategy, or confidential legal advice into an AI system unless expressly approved or directed by the City Attorney.

6. **Bias and Discrimination**: AI tools must not be used in ways that perpetuate bias, discrimination, or inequity. City users are expected to monitor these risks and take corrective action when needed.

7. **Examples**: Some examples of prohibited use include using AI as a sole basis for hiring, termination or performance evaluations, by law enforcement as a sole basis for making enforcement decisions, or by city users to benefit them personally.

Data Privacy and Security:
The following specific data privacy and security protocols apply to the use of AI:

1. **Data Handling**: City users must ensure that no city data containing personally identifiable information (PII), protected health information (PHI), or programming code is uploaded into any AI tool unless such tools are data masked or expressly approved and configured for such use by the department head and information technology department.

2. **Third-Party AI Tools**: Use of third-party AI tools should be assessed for data privacy risks. Tools should not be used without verifying that they meet city data security standards (as established by information technology (IT)). The City and its vendors shall comply with applicable privacy and security laws and contractual requirements.

3. **Security Compliance**: All AI tools must comply with the city's IT security standards, and all generated content should be stored and shared securely.

4. **Sensitive or Protected Data**: All city users of AI must remember that any sensitive, confidential or protected data (health, financial, legal) that is entered into an AI tool may become public and should refrain from entering any such data that could jeopardize the confidentiality of such information. As such, users of city AI tools shall not enter any sensitive or protected data into public facing or consumer available AI tools that are not proprietary to the city (no data ownership by the city) and/or open to public viewing. Furthermore, for AI tools that are proprietary to the city, the city must ensure that the vendor will not share any personal or confidential information entered into the AI tool, and

that there are safeguards provided to the city to ensure that there shall be no unauthorized use or viewing of such data entered by the city into the AI tool.

Ethical Considerations:
The following additional ethics related standards must be kept in mind when using AI:

1. **Transparency**: To ensure transparency, IT shall disclose through its website or other general communications, that it utilizes AI as a tool in accordance with applicable ethical standards for accountability, bias prevention, scrutiny, verification and review of any AI generated work-product. Any specific AI transparency requirements under federal or state laws shall also be adhered to.

2. **Accountability**: City users are responsible for reviewing and validating the output generated by AI systems. Any AI-generated content must be aligned with city values and standards. AI cannot replace human oversight or decision-making in all areas of work, including public safety or city regulation.

3. **Bias Prevention**: City users must ensure that AI is used in ways that do not start, perpetuate or support systemic biases.

4. **Scrutiny of Work**: When using any city approved AI tool for tasks that affect public communication, or decisions or services offered to the public, use additional steps to ensure that any final decisions that may affect any such individuals of the public are reviewed for output accuracy and avoidance of bias. Examples of such tasks include, but are not limited to, creating interview questions for job vacancies, screening materials for hiring, documents relating to regulation or safety, and any data related to policy.

Oversight and Accountability:
When using AI, the following oversight and accountability will ensure AI use is being strictly monitored:

1. **Supervisory Review**: Department supervisors shall monitor the use of approved AI tools, and any potential unauthorized use to ensure compliance with this administrative regulation. City users shall report any concerns regarding inappropriate or harmful use of AI to their supervisor or to IT.

2. **Acknowledgment of Regulation**: All AI users shall acknowledge that they have reviewed and accepted the terms and conditions of AI appropriate use as presented in this administrative regulation. A department supervisor shall ensure that this administrative regulation is distributed to all users within their department for appropriate acknowledgment and certification of review by all users.

3. **Responsibility of Information Systems**: IT is responsible for all information systems and services, including AI tools and related software. See Administrative Regulation 150. Introduction of new AI tools or pilot programs

is only permitted under a written plan approved by IT and appropriate department heads.

4. **Reporting Violations**: Any suspected misuse of AI should be reported through the City's established reporting channels, which include notification to the department involved, IT and HR.

Compliance with Laws:
In addition to this administrative regulation, all users must comply with applicable federal or state laws related to data, security, intellectual property, such as copyright and trademark laws, accessibility laws, and department required specific laws for AI use, such as California State BAR rules and regulations, human resources related laws and those related to law enforcement. Additionally, all vendor contracts with AI tool providers that involve sensitive information that may be subject to the California Consumer Privacy Act (CCPA), or the Health Insurance Portability and Accountability Act (HIPAA) should require that those vendors adhere to these laws.

California Public Record Act (CPRA):
The use of AI tools may generate content that constitutes a public record pursuant to the California Public Records Act (CPRA). A public record is any writing containing information relating to the conduct of the public's business that is prepared, owned, used, or retained by a local agency, regardless of its physical form or characteristics. Questions regarding whether an AI prompt, output, or related record must be retained, disclosed, or preserved shall be referred to the City Attorney's Office.

Violations:
Violations of this administrative regulation may result in disciplinary action, up to and including termination. Serious breaches, such as intentional misuse of AI tools to manipulate public opinion, cause harm, or breach confidentiality, will be referred to the city's human resources department, as well as to the respective department head.

Administrative Regulation Review:
This administrative regulation will be reviewed annually by the city's IT department and city attorney's office to ensure its relevance and effectiveness considering evolving technology, laws, and City needs. AI is continually changing and as such, this regulation may be amended or updated to reflect any such changes in technology as required. While this regulation covers most relevant rules pertinent to AI, it is not all inclusive, and departments may wish to create their own additional policies that supplement this administrative regulation. Reviews of this AR may also be triggered outside the regular schedule by major technology incidents or deployment of significant new AI platforms or services by the city.

Applicability of Administrative Regulation 155:
This regulation incorporates by reference Administrative Regulation 155, the Information Services Technology Use Policy, for its general provisions pertaining to technology use

by city users, including, but not limited to, ownership, retention and destruction of electronic data.

Department Specific Policies:
Each City department may also develop further internal policies and rules pertaining to appropriate AI use that are specific to their respective roles and functions and specialized uses of AI and required under federal or state laws. Any internal rules and regulations shall be reviewed by IT to ensure adherence to technical and data specifications and by the city attorney's office for adherence to the appropriate use of AI.

Last updated: 6/26/26

DEFINITIONS POLICY VERSION CONTROL ACKNOWLEDGEMENTS CONTACT US

CITY OF BOSTON GENERATIVE ARTIFICIAL INTELLIGENCE (GEN AI) POLICY

Learn more about the City's approach to artificial intelligence.

PURPOSE

Generative artificial intelligence ("Gen AI") are powerful technologies that can enable employees of the City of Boston to deliver better government services and experiences. However, Gen AI tools are not the right solution to every problem, and they create unnecessary risks and cause harm when used in the wrong situations.

This policy and its supporting resources aim to enable employees to understand and manage the risks of Gen AI tools while delivering the biggest value for our constituents. Ultimately, when these tools *support* the work of discerning and well-informed public servants, the City can get the most value while minimizing risks, delivering better services, and maintaining the trust of our constituents. The policy requires employees to complete basic training before getting access to City Gen AI tools.

QUESTIONS?

If you have questions or concerns about this policy or any Gen AI tools impacting your work, please reach out to techgovernance@boston.gov. Understanding and learning from staff experiences and perspectives helps us build a Boston that is home to everyone.

DEFINITIONS

DEFINITIONS

The following definitions can be found in the City of Boston Tech Policy Glossary. If the Glossary's "Last Update" is more recent than Mar 6, 2026 , the definitions may not match perfectly. In that case, use the definition found in the Glossary instead of this document.

Artificial Intelligence (AI): A class of technology tools that enable machines to perform complex tasks that only humans have previously been able to perform. There are different branches and types of artificial intelligence.

Generative Artificial Intelligence ("Gen AI"): A type of artificial intelligence that can learn from and mimic large amounts of data to create content such as text, images, computer code, and more, based on inputs or prompts

Data: geospatial, tabular, textual, legislative, and source code that is maintained in an electronic, digital, or optical format

City-Developed Gen AI Tool ("City-Developed Tool"): A generative AI tool that was developed and is actively maintained by City of Boston staff

City-Approved Gen AI Tool ("City-Approved Tool"): A generative AI tool that was developed by a non-City organization or company, but has been expressly approved for use by City workers

External Gen AI Tool ("External Tool"): A generative AI tool that was developed by a non-City organization or company and has not been expressly approved for City use

AI Hallucinations ("Hallucinations"): When a generative AI model outputs information that is factually incorrect or fabricated, presenting it as if it were accurate or true. For instance, a reference to a book that does not exist but that has an author, a title, a page number and other elements that make it look truthful.

POLICY

POLICY

PART 1: EMPLOYEES ARE RESPONSIBLE FOR THE IMPACTS OF GEN AI TOOLS

City employees that choose to use Gen AI tools in their work maintain ultimate responsibility for their work product and its downstream impact on constituents and City operations. They are expected to exercise human judgment and critical thinking in deciding

if, when, and how to incorporate Gen AI tools into their individual work. The use of GenAI does not absolve an employee of accountability for the accuracy, ethics, or outcomes of their assignments. The rest of this policy, the FAQ, and the Use Case Guidelines are meant to make it easier to understand the risks of using these tools in different situations, so that employees can make thoughtful and informed decisions.

Decisions about how Gen AI should fit into departmental operations are left to the discretion of departmental leadership, so long as such decisions are consistent with this policy. Individual departments may create additional policies or requirements related to Gen AI use that are more restrictive than this policy. Employees are encouraged to engage with their supervisor and department leadership to navigate expectations around work involving Gen AI.

PART 2: UNDERSTAND AND MANAGE GEN AI RISKS BEFORE USING GEN AI TOOLS

Unlike other technologies, Gen AI tools are *probabilistic*. They do not produce a predictable, determined answer based on clear inputs (like a calculator). Instead, they combine inputs with the data they were trained on to generate new content that **often cannot be predicted in advance**, which can lead to errors and hallucinations.

Depending on the context in which these errors are made, this can create significant risks for constituents and/or employees. This section clarifies what employees should do to identify and manage risks appropriately For more detail, please see the "Known Risks & Harms of Gen AI" Section in the Gen AI FAQ.

2a) Avoid Unacceptable Uses of Gen AI

There are some use cases of Gen AI that create unacceptable levels of risk. These use cases **significantly interfere with or delegate decision-making about a constituent's ability to access City services or exercise their human rights or civil liberties**.

City employees shall not use Gen AI in any of the following ways:

- To determine constituent eligibility for or access to services or benefits

- To evaluate candidates for open roles or promotions, including resume review
- To generate false, misleading, or deceptive content
- To impersonate an individual in any format without their explicit consent, including but not limited to written material, audio content, images, or video
- To translate Vital Documents (defined in City Code) without explicit consent from the Language and Communications Access Team (lca.staff@boston.gov)
- To take and/or summarize meeting notes using **External Tools** (tools that are not City-developed or City-approved)
- To take and/or summarize meeting notes using **City-Developed** or **City-Provided tools** without first obtaining informed consent from all parties.

As the field of Gen AI evolves over time, uses may be added or removed from the list above at DoIT's discretion (with clear notice to employees). City employees can read more about why these use cases are unacceptable in the FAQ.

2b) Follow Best Practices for Established Use Cases

The Use Case Guidelines provide best practices for use cases that are common to City workers. These guidelines are meant to help employees use Gen AI effectively and safely in their work. Employees interested in or actively using Gen AI are strongly encouraged to review these guidelines regularly as they are updated on an ongoing basis.

2c) Work with DoIT to Explore New Use Cases

Gen AI is a rapidly evolving field, with new tools and use cases emerging on a regular basis. Employees are expected to engage DoIT to explore new, useful applications of Gen AI in their work.

Employees are **strongly recommended** to notify DoIT if they want to use Gen AI for something that is not explicitly covered in the Use Case Guidelines. Email techgovernance@boston.gov with information about what you are trying to do with Gen AI and any tool(s) you are considering using. DoIT staff will work with you to clarify what type of risk(s) your use case poses and identify appropriate risk mitigation strategies.

PART 3: CHOOSE THE RIGHT TOOL BASED ON THE DATA YOU USE

After assessing the risks of using Gen AI for a given use case, City workers can choose from different tools to use. **This section is meant to clarify what tools are available to City workers and guide tool selection**. The full list of available City tools can be found in the City's AI Inventory.

Part 3a) Complete City Training to Access City Tools

- To access City-Developed and City-Approved tools, City workers **must** complete DoIT's approved Gen AI in Government Training, which is reviewed on an annual basis. The current version is located online.

- City workers interested in using Gen AI tools are **strongly encouraged** to complete the Responsible AI Certification through Innovate(US). More details can be found on the Gen AI Tools & Training Page in Beacon.

Part 3b) Use City-Developed Tools for Any Kind of Data

- City workers may use **City-Developed tools** to work with **any kind of data. City-Developed tools** have been developed for City workers, are actively supported by DoIT staff, and have strong security and privacy controls. They are marked as '**City-Developed**' in the AI Inventory.

- City workers are **strongly encouraged** to follow best practices in the Use Case Guidelines while using these tools.

Part 3c) Use City-Approved Tools for Medium- or Low- Sensitivity Data

- City workers may use **City-Approved tools** for **medium- and low-sensitivity data only.** City-Approved tools will have a contract in place that ensures baseline privacy and security features. They may only be available to certain employees. They are marked as 'City-Approved' in the AI Inventory.

- Employees can review the Data Security Policy for more information on what is considered **medium- and low-sensitivity data**.

- City workers are **strongly encouraged** to follow best practices in the Use Case Guidelines while using these tools.

Part 3d) Do not use External Tools for City Work

• **City workers may not use External Tools for City work**. External tools include any tool that is not explicitly listed in the AI Inventory.

ADDITIONAL RESOURCES

- Gen AI Tools and Training for Employees (Beacon Page)

- Gen AI FAQ (Beacon Page)

- Gen AI Use Case Guidelines for Employees (Beacon Page)

- City of Boston Tech Policy Glossary (Google Doc)

- Data Security Policy for Employees (Beacon Page)

DEPARTMENT-SPECIFIC RESOURCES

Some departments have opted to create additional guidance related to Generative AI that is specific to their work. DoIT does not manage these documents directly. Departments may set stricter standards than this policy, but they must meet its minimum requirements at all times.

VERSION CONTROL

DATE	VERSION	CHANGE(S)	RESPONSIBLE PERSON(S)

DATE	05/18/2023
VERSION	1.0.0
CHANGE(S)	Guidelines drafted and published
RESPONSIBLE PERSON(S)	Chief Information Officer, DoIT, Santiago Garces
DATE	01/28/2025
VERSION	2.0.0

CHANGE(S)	Complete policy revision with a focus on Enterprise-wide Gen AI tool use
RESPONSIBLE PERSON(S)	Chief Information Officer, DoIT, Santiago Garces; Director of Tech Governance & Policy, DoIT, aleja jimenez jaramillo
DATE	03/06/2026
VERSION	2.1.0
CHANGE(S)	Changes made to clarify guidance on allowable tools, relationship between policy and support documents ('supplement'), and departmental leadership discretion regarding Gen AI use
RESPONSIBLE PERSON(S)	Chief Information Officer, DoIT, Santiago Garces; Director of Tech Governance & Policy, DoIT, aleja jimenez jaramillo

ACKNOWLEDGEMENTS

The development of these guidelines has benefited from the contributions of academic, community, and City of Boston team members. Special thanks to:

- Beth Noveck, Director, Burnes Center for Social Change; Professor, Northeastern University School of Public Policy and Urban Affairs
- Dan Jackson, Director, NuLawLab, Northeastern University Law School
- John Basl, Associate Professor of Philosophy, Northeastern University Department of Philosophy & Religion
- Kade Crockford, Director, Technology and Justice Programs, ACLU of Massachusetts
- Kevin O'Reilly, Senior Manager of Innovation Programs, Howe Innovation Center at Perkins
- Kimberly D. Lucas, Professor of the Practice in Public Policy and Economic Justice, Northeastern University School of Public Policy and Urban Affairs
- Martha F. Davis, Professor, Center for Global Law & Justice, Northeastern University Law School
- Raquel Ronzone, Associate Director, Strategy & Partnerships, Howe Innovation Center at Perkins
- Sandy Lacey, Executive Director, Howe Innovation Center at Perkins
- Sarah Hubbard, Associate Director for Technology & Democracy, Allen Lab for Democracy Renovation

Still have questions? Contact:

INNOVATION AND TECHNOLOGY

617-635-4783

SEND AN EMAIL

1 CITY HALL SQUARE
ROOM 703
BOSTON, MA 02201-2021

PROVIDE YOUR FEEDBACK