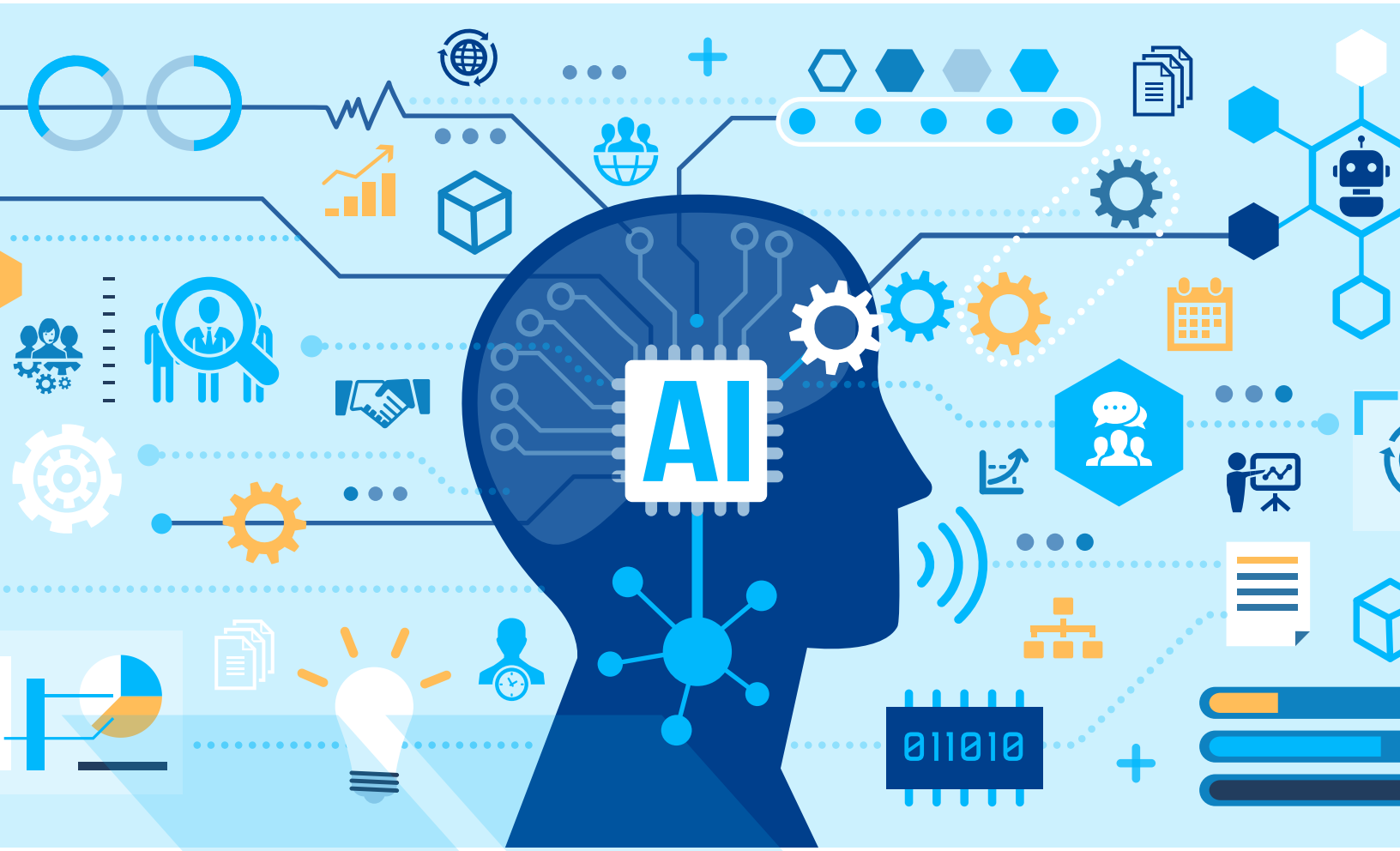
BEST PRACTICES
for AI and Workplace Assessment Technologies

UPDATED AUGUST 2026



FPF | AI

AUTHORS

Stacey Gray, Senior Director for Artificial Intelligence
Tatiana Rice, Senior Director for U.S. Legislation

CONTRIBUTORS

The authors would like to thank the members of the Expert Working Group who contributed to this resource, including experts from Dayforce, LinkedIn, UKG, and Workday. In addition, Future of Privacy Forum staff who supported, edited, or offered feedback include James Hamraie, Lauren Nieto, Matthew Reisman, Josh Lee Kok Thong, and Vincenzo Tiani.

ACKNOWLEDGEMENTS

Special thanks to Brenda Leong (Director, AI Division, Zwillgen) and Nicolette Nowak (General Counsel, Beamery) for their thoughtful and helpful feedback on earlier drafts of this publication.

FPF | CENTER FOR AI

The Future of Privacy Forum (FPF) is a global non-profit organization that advances principled and pragmatic data protection, AI and digital governance practices. We convene leaders across industry, academia, and the public sector to provide expert analysis, benchmarking, and best practices that support responsible innovation and regulatory compliance. Learn more about FPF by visiting fpf.org.

The **Center for Artificial Intelligence** at the Future of Privacy Forum is dedicated to navigating the complex landscape of AI governance and its intersection with privacy and data protection law. Drawing on expertise from a global Leadership Council comprising industry leaders, academics, civil society, and policymakers, the Center provides sophisticated, practical policy analysis to help organizations align innovation with responsible implementation while meeting evolving regulatory requirements. Learn more about the FPF Center for AI at fpf.org/ai.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Since the Future of Privacy Forum published its Best Practices for AI and Workplace Assessment Technologies in 2023, the use of artificial intelligence across the employment lifecycle has expanded considerably, and the underlying technology has changed. The 2023 guidance addressed predominantly predictive systems used in discrete decisions such as recruiting, hiring, and promotion. Employers today also deploy generative and agentic systems that draft communications, summarize candidate information, conduct conversational interviews, and are capable of executing multi-step workflows with limited human intervention. These capabilities can support efficiency, widen the range of candidates employers consider, promote more consistent decision criteria, and benefit job seekers and employees themselves, but also carry risks that earlier guidance did not anticipate, including: non-deterministic outputs, hallucinated content, reduced auditability, and the capacity to act directly on data and connected systems. The legal and governance environment has also grown more complex, as new laws and maturing voluntary frameworks shape responsible development and deployment expectations.

The Updated Best Practices retain the commitments that anchored the original framework: responsible AI governance, non-discrimination, transparency, data security and privacy, and human oversight. As a resource for practitioners and policymakers, this updated version also begins with two foundational concepts:

- **Understanding the AI Value Chain** — distinguishing the roles and degrees of control held by foundation model providers, developers, deployers, and end users. In the context of employment systems, primary responsibilities for AI governance lie with developers and deployers.
- **Assessing Risk** — providing a framework for assessing the risk of a particular employment use: the quality and sensitivity of the data and inferences involved, the degree of independent judgment the system exercises, the proximity of its output to the final decision, and the significance of that decision for the individual or employer organization. Risk in this framing is contextual, and depends on both design and configuration, and organizations can scale the application of best practices in accordance with the level of risk.

The Updated Best Practices that follow (in Part III) assign responsibilities to Developers and Deployers across that spectrum:

- Responsible AI governance **(Part III-1)** calls for programs that govern, map, measure, and manage risk across the system lifecycle, consistent with the functions of the NIST AI Risk Management Framework, and accounts separately for predictive, generative, and agentic behavior.
- The non-discrimination practices **(Part III-2)** direct organizations to comply with applicable anti-discrimination law and to test proactively for unintended bias, using methods such as counterfactual and shadow testing, and proxy techniques where demographic data is limited.
- The transparency practices **(Part III-3)** divide disclosure duties between Developers, who document system characteristics and intended uses, and Deployers, who provide accessible and accurate notice to affected individuals.
- The data security and privacy practices **(Part III-4)** address safeguards for personal data alongside newer exposures such as p rompt injection, data leakage, and agentic access to connected systems.
- Finally, the human oversight practices **(Part III-5)** describe a graduated set of authority levels, from full human control to bounded autonomous action, and hold that humans remain accountable for consequential outcomes regardless of the authority delegated to a system.

These Updated Best Practices are intended to serve as a holistic practitioner framework, and draw throughout on the NIST AI Risk Management Framework and related NIST guidance, ISO/IEC 42001 and 42005, relevant provisions of the EU AI Act, and emerging U.S. state frameworks, including those in California, Colorado, and Connecticut. They are designed to help Developers and Deployers meet current obligations under civil rights, employment, and privacy law while preparing for a governance environment that continues to evolve.

I. Introduction

Since the Future of Privacy Forum published its "Best Practices for AI and Workplace Assessment Technologies" in 2023, the integration of artificial intelligence (AI) systems in employment settings has accelerated dramatically. The capabilities of AI systems themselves have transformed: from predominantly predictive, decision-support tools toward generative AI and agentic systems capable of autonomous action across the employment lifecycle.

These developments bring significant new opportunities, but also new and compounding risks — particularly for vulnerable or underrepresented communities. Employment decisions exist at the heart of individual dignity and opportunity, governing access to livelihood, economic security, and professional standing, and they are often made within relationships marked by significant asymmetries of information and bargaining power between workers and employers. Against that backdrop, they are governed by a wide range of long-standing employment and civil rights laws. In recent years, the legal, regulatory, and governance landscape has grown substantially more complex, intensifying the challenge of managing these systems responsibly.

In 2023, the Best Practices primarily addressed the use of predictive AI systems in consequential employment decisions, such as recruiting, hiring, promotion, and termination.[1] In this Updated Best Practices, the consequential decisions in scope have not fundamentally changed, although today's AI systems are now deployed more broadly — in recruiting, hiring, performance evaluation, compensation analysis, and workforce management. What has changed is the AI itself: the field has expanded beyond predictive AI systems to include **generative and agentic systems** that can draft, communicate, recommend, and increasingly *act* across the employment lifecycle with varying degrees of autonomy and human oversight.

Generative AI and agentic AI systems have a range of potential applications. In contrast to predictive systems that *score or rank* inputs based on historical data, a generative AI system can produce *new content and communicate in natural language* — for example, to draft a job description, summarize candidates' information, or conduct conversational interviews. Agentic systems can go further, to execute multi-step workflows, such as an autonomous recruiting agent that searches candidate pools, conducts outreach, schedules interviews, and screens, scores, or ranks applicants. Both introduce new risks related to hallucination, security, reduced transparency and auditability, and the exposure of sensitive personal data, all of which complicate existing governance approaches.[2]

Meanwhile, the legal and voluntary framework landscape informing AI in employment has evolved since 2023, and shapes how risk is understood. Jurisdictions around the world have enacted laws and regulations on AI in employment, including through privacy and data protection laws that cover automated decision-making and profiling.[3] Voluntary frameworks and technical standards, such as ISO/IEC 42001, NIST AI RMF, and sector-specific initiatives, have matured in ways that further inform what responsible AI development and deployment requires.[4] Although law and policy frameworks often treat employment as categorically "high risk," the practical reality is that risk is often shaped by the specifics of context and use, with some employment uses of AI being relatively lower risk (like summarization or job description generation), and others requiring much greater oversight and governance (like assessing promotion readiness or evaluating interview performance). Accordingly, these Updated Best Practices begin with a Foundational Concepts section that offers a novel **risk framework** for assessing specific uses of AI in employment, recognizing that higher risk uses warrant more rigorous governance.

Amid all that is changing in AI, core principles for responsible governance remain the same. Developers and deployers of AI systems that affect employment have a responsibility to build and use them in a manner that complies with civil rights, employment, privacy, and AI laws, and is consistent with consensus best practices, standards, and frameworks. These Updated Best Practices are designed to help organizations meet that responsibility — navigating a more complex governance environment while maintaining the core commitments to responsible governance, non-discrimination, transparency, data security and privacy, and human oversight that have anchored this framework from the outset.

II. Foundational Concepts: Roles, Risk, and Scope

Effective AI governance depends on a shared foundation: who bears responsibility, what risks are present, which practices apply and under what conditions. With that foundation, organizations can operationalize best practices in ways that meaningfully address the risks they are designed to mitigate. This clarity is especially critical in the employment context, where AI systems may inform consequential decisions that impact individual livelihoods and responsibility for those systems is distributed across a complex chain of actors with varying degrees of control.

This Part develops that foundation across two areas. The first maps how governance responsibility is distributed across the actors who develop, deploy, and use employment-related AI systems — recognizing that each occupies a distinct position in the value chain with different degrees of control, visibility into system behavior, and capacity to mitigate risk. The second provides a framework for assessing and classifying risk, identifying the factors that elevate certain AI tools to high-risk status, and informing when heightened governance measures in this guide are warranted. Together, these foundational concepts can guide organizations towards determining whether and to what degree the best practices detailed in the following sections regarding AI governance, non-discrimination, transparency, data security and privacy, and human oversight should apply.

A. UNDERSTANDING THE AI VALUE CHAIN

AI systems that impact employment opportunities typically flow through a complex value chain involving a range of actors — foundation model developers, AI system developers (often vendors), deployers (often employers), and end users such as HR professionals and hiring managers. Each maintains different degrees of control and visibility into system behavior, and has different capacity to mitigate risks. Governance responsibility is therefore both shared and differentiated: upstream actors set constraints on model capabilities and determine system design, while downstream actors determine how systems are configured, deployed, and used in practice.

The **Updated Best Practices in this resource (Part III)** primarily describe the responsibilities of AI System Developers (“Developers”) and AI System Deployers (“Deployers”), but are informed by an understanding that responsibilities and accountability often flow upstream (to general-purpose model developers) or downstream (to end users).

TYPICAL ACTORS IN THE AI VALUE CHAIN	
Foundation Model Providers	**Develop and provide general purpose AI models — distributed either as open-source or via SaaS APIs — trained on massive, diverse datasets.** They are responsible for fundamental design choices, including training data selection, model architecture, pre-deployment safety testing, and capability boundaries.
Developers of AI Systems	**Vendors who build AI systems by adapting foundation models or creating task-specific models into product offerings for particular employment applications** — responsible for translating AI models into systems designed for employment use cases; often control fine-tuning, guardrails, system architecture, application-specific testing, and documentation of system capabilities and limitations.
Deployers of AI Systems	**Employers or operators who configure and implement AI systems into HR processes** — responsible for maintaining ongoing control over real-time system functions and operational conditions and applications, transparency to affected individuals, and human oversight implementation.
End Users	**Personnel who interact directly with AI systems** — typically responsible for using systems in line with policies and providing feedback about real-world performance; end users can support the deployer in assessing whether systems align with the organizational compliance policies and strategy.

Governance responsibility in the employment AI value chain flows from development to deployment and use, and back through feedback loops that shape how systems are refined and governed over time.

B. UNDERSTANDING HIGH RISK: A FOUR-FACTOR RISK FRAMEWORK

Risk in the employment context is not a fixed property of an AI system. It is context-specific and emerges from the interaction of a system's intended use and capabilities, the data it draws on, how directly its output drives a decision, and the importance or stakes of that decision. The same underlying model or system may present very different risks depending on how and where it is configured or deployed. For example, an AI system used by employers to rank job applicants directly shapes access to employment, and may thus pose a heightened risk. In contrast, a candidate-facing tool that recommends job opportunities based on current vacancies and candidate-provided information may present lower risks.[5]

For these reasons, most emerging AI laws approach the concept of "high risk" in part based on intended use cases and key applications.[6] However, the binary approach of high vs. low risk categories can be limited. Many employment use cases are not necessarily high risk (e.g., a scheduling assistant, or informational chatbot for HR-related questions), or only become "high risk" when deployed or configured in certain ways. At the same time, many "low risk" use cases can nonetheless contribute to, inform, or shape, significant decisionmaking down the line — and therefore may deserve heightened governance protections.

In order to shine a light on this complexity, we provide a four-factor risk framework for practitioners to assess and classify an AI system's overall risk profile in the context of hiring and employment. The purpose of this broader analysis is to allow organizations to calibrate their governance measures to the obligations in the Updated Best Practices below (Part III), with more rigorous implementation for higher risk systems and contexts.

Relevant risk factors include:

- **Quality and Sensitivity of Data and Inferences (Input and Output)**
- **Degree of Autonomy, Discretion, and Action (Decisioning)**
- **Proximity to the Decision (Output / How Output Is Used)**
- **Nature and Significance of Impact (Consequentiality)**

No single factor is determinative. To determine risk classification, organizations should assess where their AI system falls along a spectrum for each risk factor, considering how factors combine to elevate or reduce overall risk, as well as how each dimension may carry associated legal standards and compliance considerations. These considerations are relevant for both developers, who may determine the risk level of their offerings based on their intended use and deployment context, and deployers, whose risk classification should draw on such developer determinations and their own configuration, deployment context, and use of the AI system.

FACTOR 1: Quality and Sensitivity of Data and Inferences (Input & Output)

The core questions for this risk factor are: ***What data is used in training or fine-tuning a model? What data is used as system input, does the system draw on, and what predictions or inferences does it generate in system output?***

KEY CONSIDERATIONS

- **Sources** — Was the data collected directly from the individual (e.g., application materials, self-reported information)? If not, where did it originate — employer records, third-party sources, publicly available data? Does the data include sensitive data?
- **Relevance and Accuracy** — Is the data complete, accurate and directly relevant to the consequential employment decision at hand?
- **Inferences** — Do the inferences originating from the system touch on protected characteristics or serve as proxies for them? Do those inferences touch on any sensitive or protected characteristic or status (e.g., an emotional state, a health condition, or predicted behavior)?[7]
- **Expectations** — Would the individual reasonably expect this data to be used in an employment context, or in this specific employment context? Would the individual expect particular inferences to be made?

LOWER RISK EXAMPLES

HIGHER RISK EXAMPLES

LOWER RISK EXAMPLES		HIGHER RISK EXAMPLES
An AI system summarizes directly provided, job-relevant data (resumes, performance records).	**Sources and Relevance (INPUT)**	*An AI system summarizes third-party data, not directly related to the requirements of the job.*
A generative AI system processes employee writing (input) and provides the employee with suggested improvements to grammar or persuasiveness.	**Inferences and Expectations (OUTPUT)**	*A generative AI system processes employee writing and infers the employee's confidence, anxiety, or enthusiasm, flagging for supervisors as "low engagement" or "high stress" (likely to be unexpected inferences that affect protected classes)*

FACTOR 2: Degree of Autonomy, Discretion, and Action

The core question for this risk factor is: ***To what extent does the AI system exercise independent judgment or discretion in reaching its output?***

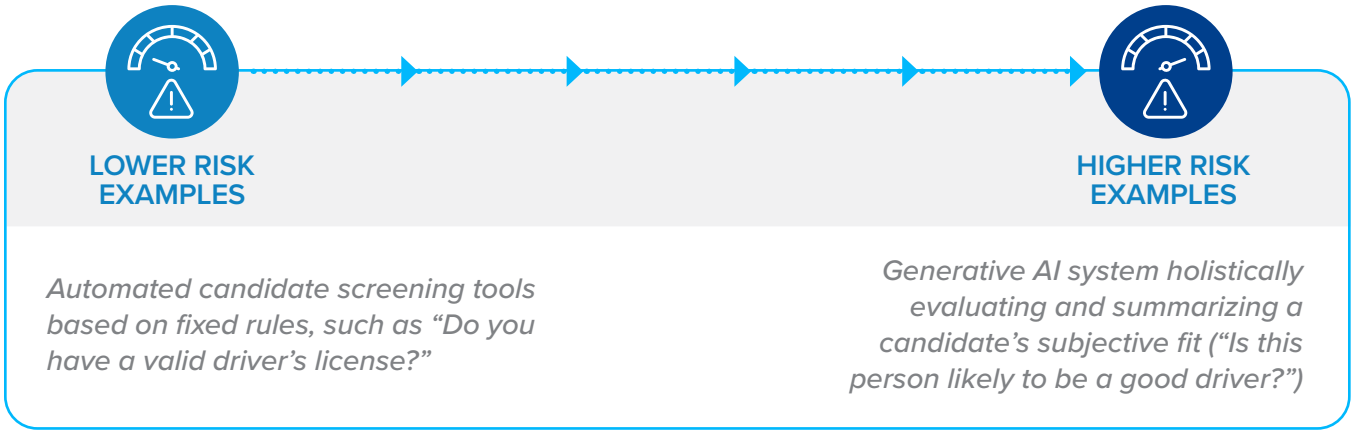
KEY CONSIDERATIONS

- **Decision Logic and Latitude —** Is the system rules-based, deterministic (e.g., traditional machine learning), or probabilistic e.g., generative or agentic AI)? To what degree does the system create or apply autonomous reasoning to evaluate inputs and reach an output?
- **Transparency —** Can the system's reasoning or outputs be traced and explained, or is it opaque?
- **Scope of Action —** Is the system bound to a single task, or does it operate autonomously across multiple steps?
- **Output Constraints —** Are there limits on the range of outputs the system can produce?

Not all AI systems exercise the same degree of autonomous reasoning. Rules-based automated systems operate on fixed, predetermined logic — if X, then Y. Risk is bounded by the rules themselves, making these systems more auditable and predictable, even though the rules themselves may encode bias or error. In contrast, machine learning models are trained on large datasets to detect patterns and generate predictions or scores. Reasoning can be less transparent and more difficult to trace, and the system may perpetuate or amplify patterns present in training data.

In recent years, large language models (LLMs) have introduced more novel considerations. Their reliance on probabilistic reasoning means that identical inputs will not always yield the same output. LLM-based systems are designed to apply broad general-purpose reasoning to interpret inputs, weigh factors, and generate outputs — including narrative evaluations, recommendations, or assessments — making their outputs harder to predict, audit, or contest.

Finally, **agentic AI systems**, using protocols that integrate large language models, can operate autonomously across multiple steps, taking actions and making sequential decisions with limited human intervention. In employment contexts, this may include systems that independently source candidates, schedule interactions, or execute workflows — compounding the risk of any single error across a chain of actions.[8] Agentic workflows can also exercise discretion in how to accomplish tasks.



FACTOR 3: Proximity to the Decision (Output / How Output Is Used)

The core question for this risk factor is: ***How close is the system's output to the final decision?***

KEY CONSIDERATIONS

- **Purpose of Use —** If used for decisionmaking, does the system's output feed directly into a final decision, or does it sit earlier in the decision-making process?
- **Influence of Output —** If used for decisionmaking, is the system's output one input among many, or does it directly drive a decision? If the former, how much weight does it carry relative to other factors informing the final decision?[9]
- **Human Review —** Is there meaningful human review of the system's output before it is used for a decision?
- **Reversibility —** If the system's output contains an error, can it be identified and corrected, or does the output foreclose options before a human is involved?

LOWER RISK EXAMPLES

AI chatbot that answers questions about internal HR policies, such as vacation days or benefits.

HIGHER RISK EXAMPLES

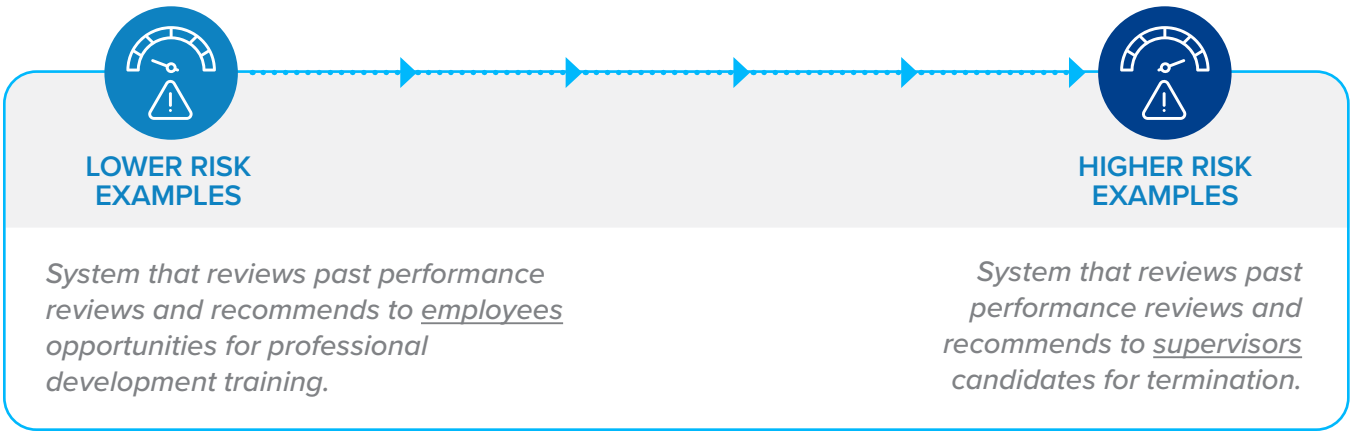
AI chatbot with agentic capabilities that responds to, and independently approves, requests for time off or other internal HR benefits.

FACTOR 4: Nature and Significance of Impact (Consequentiality)

The core question for this risk factor is: ***What kind of decision does the system inform, and how significant are the consequences for the individual?***

KEY CONSIDERATIONS

- **Type of Decision —** What category of employment decision does the system inform, and what legal rights and protections attach to that type of decision?
- **Material Impact —** How significantly does the outcome affect the individual's livelihood, earnings, career trajectory, or working conditions?
- **Severity and Reversibility —** How serious is the potential harm if the system produces an erroneous or biased output? Can it be identified and corrected, or does the output foreclose options before a human is involved?
- **Breadth of Impact —** Does the system's output affect one individual or many — and could it produce systematic disparate impact across protected classes?



Determining Overall Risk Level

Applying this framework in practice means treating each factor as a discrete lens before arriving at a composite risk determination — recognizing that risk emerges from the interaction across all four dimensions, not any single one alone. A system drawing on sensitive data may present limited risk if it operates deterministically, sits far from a final decision, and informs only administrative functions. Conversely, a system with minimal data sensitivity may present significant risk if it operates autonomously, drives consequential employment decisions directly, and affords little meaningful human review. Organizations should assess each factor independently before synthesizing across them, whether through a formal scoring approach that weights and aggregates factor-level assessments, a qualitative determination that identifies the highest-risk dimension as the ceiling, or a threshold approach that triggers escalation when any single factor reaches a critical level.

The Updated Best Practices that follow are designed for AI systems that present heightened risk profiles under this framework, and the degree of rigor with which they should be implemented can scale with the overall risk level an organization determines.

III. Updated Best Practices for AI and Workplace Assessment Technologies

The Updated Best Practices that follow are organized around the same commitments that anchored the 2023 Best Practices — responsible AI governance, non-discrimination, transparency, data security and privacy, and human oversight. Each section incorporates new considerations for generative and agentic systems where their behavior — non-deterministic outputs, fabricated or unsupported content, reduced auditability, and the capacity to act — alters how a given practice must be carried out.

These Best Practices are informed by the NIST AI Risk Management Framework and related NIST AI guidance, ISO/IEC AI standards, relevant provisions of the EU AI Act, and other applicable AI governance, privacy, cybersecurity, employment, and anti-discrimination laws, regulations, and guidance, including emerging frameworks in California and Colorado.

Specifically, these practices set out the responsibilities of Developers and Deployers, the two actors with the most direct and continuing control over how employment-related AI Systems are built, configured, and used. As Part II describes, responsibility extends beyond these two roles: foundation model providers upstream shape capabilities developers inherit, and end users downstream may determine how a system operates in practice. The rigor with which each practice should be implemented is not uniform. It is intended to scale with the risk profile established under the four-factor framework in Part II — higher-risk systems and contexts warrant more demanding application of the same underlying practices.

KEY TERMS

- **Developer** — An entity that designs, codes, creates, or modifies an AI system, including by adapting foundation models, whether for internal use or for use by third parties.
- **Deployer** — An entity that implements an AI System, including making it available to end users such as their HR team members, managers and/or employees.
- **AI System**[10] — An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI Systems vary in their levels of autonomy and adaptiveness after deployment.

  *Note: In the context of these Best Practices, the term "AI System" refers primarily to AI Systems that are **used in the context of hiring and employment and present heightened risk profiles**. Risks can be assessed using Part II and informed by any relevant legal standards. For developers, AI Systems may include both customer-facing AI Systems as well as their own AI Systems built or procured for internal use.*
- **Generative AI** — Generative AI refers to the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content — text, images, audio, video, or code — as distinct from predictive systems that score or rank inputs against historical data.[11]
- **Agentic AI System**[12]— An AI System that can plan and execute multi-step actions toward a goal, including by interacting with other systems, with limited human intervention between steps.

1. RESPONSIBLE AI GOVERNANCE

Institute AI Governance Practices to Govern, Map, Measure, and Manage Risks Across the AI Lifecycle

BASELINE PRINCIPLES

- Developers and Deployers should each maintain responsible AI governance programs, with responsibilities apportioned according to each actor's role, degree of control, and ability to identify and mitigate risks.
- Organizations should govern, map, measure, and manage reasonably foreseeable risks from design through deprecation, with the rigor of controls scaling to an AI System's risk profile.
- Governance should be anchored in clearly articulated responsible AI principles — such as fairness, accountability, transparency, safety, security, privacy, and human oversight — that are operationalized through concrete roles, processes, and accountability tools, and revisited as AI capabilities evolve across predictive, generative, and agentic functionality.
- *Obligations in this section are informed by the NIST AI Risk Management Framework, as well as ISO/IEC 42001, ISO/IEC 42005, relevant provisions of the EU AI Act, and other relevant AI governance laws, regulations, and guidance.*

1.1 Shared Responsibilities

Developers and Deployers of AI Systems should institute responsible AI governance programs to govern, map, measure, and manage risks across the AI System lifecycle.[13] Responsibilities should be apportioned between the actors to the extent of their control and ability to mitigate the risks.

Common Responsibilities of Developers

- Mapping reasonably foreseeable potential risks in context of the expected use cases and user base, including, but not limited to, risks to individuals, the organization, communities, and society more broadly;
- Documenting risks and designing mitigations, accounting for risks such as the potential for historical, sampling, and labeling biases in datasets. This includes any bias in the training data sets they use, especially in which the data sets use methods that attribute diverse characteristics (as compared to self-identification of individuals);
- Understanding and documenting limitations of their AI Systems;
- Making clear the intended user population (e.g., trained HR professionals versus front-line workers), and intended use case and purposes so Deployers can set appropriate access, training, and acceptable use policies;
- Providing configuration guidance to Deployer organizations and the administrators responsible for configuring and operating the AI System (e.g., HR, IT, or functional system administrators) to support successful deployment, including documented limitations and expected use cases; and
- Implementing security, logging, and testing capabilities that enable deployers to exercise meaningful oversight.

General Responsibilities of Deployers

- Assessing and evaluating an AI System in the context of their use cases and workforce;
- Mapping reasonably foreseeable potential risks of the AI System being used in the context of their HR processes and expected user base, including, but not limited to, risks to individuals, the organization, communities and society;
- Where possible, configuring the system, including access and human oversight (such as human-in-loop, if appropriate), consistent with intended uses, risks, and documented limitations of the AI System;
- Setting policies for appropriate and expected uses of an AI System, including reviewing, on an appropriate cadence, the settings and controls provided by the Developer and configured by the Deployer;
- Providing appropriate training to intended end users of an AI System (e.g. recruiters, HR professionals, hiring managers, or employees), to ensure that uses remain consistent with the Deployer's organizational policies, risk tolerances, and applicable law; and
- Monitoring AI System operation for compliance with legal obligations and remediating adverse impacts where appropriate.

1.2 Responsible AI Governance Programs

A responsible AI governance program is the organizational structure that defines the roles, processes, and oversight that allow an organization to govern, map, measure, and manage AI risks across the lifecycle.

AI Governance Programs (Generally)

Developers and Deployers should each establish internal responsible AI governance programs that take account of the risks of an AI System and context for its use. Programs should scale according to risks, with increased governance and stricter controls for higher risk AI Systems (See **Part II – Risk Factors**).

Governance frameworks should:

- Establish and document governance processes that **map, measure, and manage** reasonably foreseeable risks and harms to individuals that AI Systems can pose, particularly those that have heightened risk profiles;
- Address the full AI lifecycle (design, development, testing, deployment, monitoring, update, and retirement), recognizing the different roles and responsibilities of the actors across the AI value chain, and accounting for information and technical limitations of the different actors;
- Include an accountability and oversight structure that highlights potential risks and routes such risks for analysis and potential mitigation, along with on-going oversight and enforcement of internal policies and controls;
- Operationalize ongoing feedback and review throughout the AI System's lifecycle. Feedback should include appropriate and relevant internal stakeholders, such as members of corporate leadership, product and engineering teams, privacy and security teams, AI governance practitioners; technical and legal experts; and HR professionals with day-to-day responsibility for hiring, promotion, and termination processes; and
- Distinguish among predictive, generative, and agentic functionality when designing controls, recognizing that these increasingly coexist within the same systems and that each may carry a different risk profile — and therefore governance should account for each individually and for their combined effect.

Internal Structures and Roles

Responsible AI governance programs generally involve cross-functional participation from a broad range of internal stakeholders who meet regularly.

Internal AI governance structures should:

- Establish clearly defined internal roles and assign accountability for risk identification, mitigation, approval, monitoring, and other decision-making, in accordance with different areas of expertise, control, and abilities to mitigate risks;
- Ensure AI Systems are fit for purpose;
- Define escalation pathways for higher risk AI Systems or new and emerging risks that arise after development or deployment;
- Establish documentation and recordkeeping standards and ownership; and
- Provide mechanisms for periodic review and continuous improvement, including revisiting control design as new capabilities in AI models and systems emerge over time.

EXAMPLE

A typical AI Governance Team may include participants from product and engineering, privacy, data governance, cybersecurity, HR operations, compliance, employment legal and regulatory, enterprise risk and audit, and ethics teams.

1.3 Specific Accountability Mechanisms and Tools

Responsible AI governance programs are operationalized through internal mechanisms for risk management, monitoring, escalation, testing, and reporting. This ensures that there is a team or individual who verifies each of the key internal mechanisms are in effect.

Risk Management Processes

Developers and Deployers should have documented risk management processes that identify intended use and reasonably foreseeable misuse and assess impact on individuals — whether job candidates, employees or the broader organization — and identify and implement risk mitigation strategies accordingly. Depending on the organizations' role as either a Developer or Deployer in the AI Value Chain (**Part II**), risk management processes should also evaluate performance, reliability and fairness of the AI System and monitor outcomes over time. This includes adjusting risk mitigation strategies as needed. They should be able to consult, where appropriate, with outside experts to help identify risks and help design risk mitigation strategies.

- ***Generative AI*** requires distinct governance attention because it is non-deterministic; risk management processes should account for output variability, the potential for fabricated or unsupported content (hallucination), and adversarial inputs designed to manipulate outputs (prompt injection).
- ***Agentic systems*** that execute actions or interface with other systems introduce additional operational and security risks. Governance should explicitly account for which internal and external systems an agent can access, what each connection exposes, the controls against prompt injection

and data leakage that those connections require, and where human oversight and approval sit relative to the actions the agent can take.[14] ***(See Section 5, Human Oversight and Agentic Authority Levels.)***

Impact Assessments

Developers and Deployers should conduct **impact assessments** for their AI Systems to identify, document, and mitigate risks — with more rigorous impact assessments for AI Systems with heightened risk profiles.[15] Impact assessments should document proposed mitigation strategies for identified risks, and whether those proportionate mitigations were adopted. Assessments should be "living documents" that are revisited on a regular basis or as new risks arise, which may reflect changes in system functionality, expanded uses, or results of internal or external audits.

Impact assessments will also differ by role in the AI value chain:

- For Developers, impact assessments should be more focused on the intended uses of the AI System, model and design risks, and may include considerations such as data used to train the AI System, learning approaches, and the model architecture. Assessments should address risks related to discrimination, robustness, adversarial misuse, data leakage, hallucination risks, and agentic failure modes.[16]
- For Deployers, assessments should tend to be more focused on the context of use and include considerations on how the tool will be used within broader employment processes, the expected user base, whether outputs will influence consequential decisions, the adequacy of human oversight, and assessment of risks of unlawful discrimination. For Generative AI tools, Deployers should also assess whether output variability could change the substantive basis for a decision, and whether outputs are being relied on beyond their validated use.

EXAMPLE

An impact assessment for an AI system that sorts and ranks job candidate resumes should be more rigorous than an impact assessment for a lower risk AI system, such as one that is limited to allowing an HR Team to use natural language (rather than SQL) to query an existing database in similar ways.

Data Governance

- Developers should establish practices and procedures to ensure data used to train or fine-tune models is appropriate and sufficiently diverse and representative in consideration of the tool's intended use.
- When training a model themselves, Developers should define standards for what data should or should not be used to train or fine-tune an AI System, taking into account risks such as proxy bias, data provenance, labeling processes (as relevant), and known limitations.
- Developers and Deployers should apply data minimization, purpose limitation, retention and deletion practices;
- Developers should monitor for drift;[17]
- Developers and Deployers should take account of secure handling and appropriate access segmentation of sensitive HR data and controls for system integrations and agentic access.

EXAMPLE

A deployer connects an AI scheduling agent to its applicant-tracking system and email so it can arrange interviews. Because the agent can both read applicant records and send messages on the organization's behalf, data governance should limit the agent's access to only the records it needs, log the actions it takes, and prevent it from exposing one candidate's information in communications with another.

Evaluations, Testing, and Monitoring

Developers and Deployers should implement internal evaluations or other internal compliance structures that allow for independent feedback mechanisms. These may include:

- Structures that allow internal team members to escalate an AI System's potential risks or harms;[18]
- Methodologies for testing (or sandbox), validating, and recommending corrections to AI Systems both pre-deployment, and as well as for monitoring and re-testing post-deployment. These should include evaluation criteria for bias, risk, and accuracy.
- Testing and monitoring methodologies that are appropriate to system functionality, including generative, predictive and agentic AI Systems. For Generative AI tools, testing should account for variability across runs, and should evaluate for fabricated content and for susceptibility to adversarial or manipulated inputs (for example, through red-teaming). For agentic tools, testing should extend to the actions the system can take and the systems it can reach, not only the content it produces.[19]

Continuous Monitoring and Feedback Channels

- Both Developers and Deployers should maintain continuous monitoring and structured feedback channels. Monitoring and feedback mechanisms should include internal reporting for unexpected or harmful outputs, governance-body reviews, defined escalation criteria and remediation procedures, and documentation of decisions and corrective actions.
- Where appropriate, organizations should provide mechanisms to collect input from affected stakeholders (e.g., workers, hiring managers), while recognizing practical and legal constraints on sharing proprietary design details. Monitoring should also evaluate whether oversight controls (including human-approval workflows, logging, accountability for users, and access restrictions) remain effective as systems evolve.

EXAMPLE

A deployer's HR team notices that a resume-screening tool surfaces fewer candidates from a particular regional talent pool after a model update. Under a defined escalation pathway, the team routes the anomaly to the organization's AI governance committee, which can evaluate whether remediations are necessary, and if so, decide to pause the tool, request assistance from the developer, and document the decision.

2. NON-DISCRIMINATION

Organizations Should Test for Unintended Bias and Comply with Non-Discrimination Laws

BASELINE PRINCIPLES

- AI Systems can introduce or amplify societal biases that can result in unlawful discrimination when deployed in higher risk contexts.[20] At the same time, AI Systems can be a powerful mechanism to *limit* discrimination resulting from human biases, including by enabling more standardized treatment of similarly qualified candidates or employees.[21]
- Both Developers and Deployers have obligations to comply with a wide range of global anti-discrimination laws, and should proactively implement governance frameworks to account for future laws and regulations.
- Even when not legally required, Developers and Deployers should proactively engage in self-testing for unintended bias and unlawful discrimination, including by taking into account evolving testing standards and practices as part of a broader AI Governance Program (**Section 1**).
- Developers and Deployers differ in their access to training data and visibility into real-world applications. As a result, Developers are best positioned to test a system and articulate its intended uses, while Deployers are best positioned to assess the fitness for their specific use cases.
- *Obligations in this section are informed by the NIST AI Risk Management Framework and related guidance, as well as ISO/IEC 42001, relevant provisions of the EU AI Act, and other relevant anti-discrimination, employment, and AI governance laws, regulations, and guidance.*

2.1 Legal Compliance

Developers and Deployers share a broad set of anti-discrimination legal obligations, but the specific requirements and allocation of responsibilities vary by jurisdiction and use case. The principles below apply to both parties, with role-specific responsibilities developed in Section 2.2.

- **Legal compliance.** Developers and Deployers must comply with a range of federal, state, and global anti-discrimination laws, and should proactively implement governance frameworks that can adapt to future laws and regulations.
- **Proactive self-testing.** Even when not legally required, Developers and Deployers should proactively engage in self-testing for unintended bias and potential discrimination, including by taking into account evolving testing standards and practices as part of a broader AI Governance Program (**Section 1**).

2.2 Shared Responsibilities

- **Reviewing available information from model providers.** Both Developers and Deployers should review information that is available to them from the provider of the AI model and/or AI System, as applicable. For developers of AI Systems, this may include information from upstream general-purpose AI model developers (such as model cards). Information may include the steps the provider took to ensure representative data sets were used in training a model and/or any explanations of testing methodologies and results that are available.
- **Information sharing and cooperation.** Developers and Deployers should reasonably cooperate to share information for assessing and testing of AI Systems that have heightened risk profiles, with such responsibilities addressed in their agreements.

EXAMPLE

A developer of an agentic system that is intended to substantially shape and inform hiring decisions provides deployers with information they need to conduct impact assessments. The developer may also agree to provide a deployer with "test access" to the system so that it can be tested for the deployer's specific use cases, using the deployer's own hiring data, or synthetic data provided by the developer.

- **Assessing fitness for purpose.** AI Systems can have significant unanticipated impacts when they are used for unintended or secondary purposes. Identifying the intended purpose is important to ground testing in a particular context and increase its effectiveness.
  - **Developers should** clearly articulate the intended purpose for their AI Systems, including where they are intended to be used for higher risk or consequential decisionmaking.
  - **Deployers should** determine whether the AI System carries a heightened risk profile (see Part II) for their use cases, and ensure the AI System is deployed consistent with its stated purposes and consistent with existing legal obligations. If a Deployer determines that their implementation and use of the AI System is likely to raise additional risks of unintentional bias that can result in unlawful discrimination, Deployers should reconsider whether their implementation of the AI System is fit for the intended purpose, aligned with current legal standards, and additional risks are addressable.

- **Self-Testing.** Developers and Deployers should, to the extent reasonably feasible, engage in internal testing of their AI Systems for unintended bias that may cause unlawful discrimination within intended use cases related to higher risk or consequential decision-making. The testing methodologies should be tailored to the nature of the AI System and the datasets that are available to the tester.
  - **Developers should**, to the extent reasonably feasible, test their AI Systems for unintended bias against the systems' intended use cases, and ensure that the datasets used to train an AI System are appropriately representative of the population where the AI System is intended to be used. Because Developers are often technically and contractually restricted from accessing a Deployer's data, Developer testing generally relies on aggregated, publicly available, or synthetic data that simulates real-world conditions. Developers nonetheless have unique insight into model and system design choices that can guide effective testing strategies.

- **Deployers should** tailor their testing to their actual implementation and use of the AI System, scaled to the size of the deployment, the sensitivity of the decision, and the resources available, consistent with their broader AI governance program. They should seek appropriate transparency or contractual assurances from developers regarding bias testing. In other cases, a Deployer may choose to test the AI System using their own data or synthetic data, and/or implement quality controls when humans perform the same tasks as the AI System and the results can be compared.

EXAMPLES OF TESTING TECHNIQUES

Red teaming; counterfactual testing (varying protected or proxy characteristics to observe whether outputs change); disparate-impact analysis (including statistical screens such as the four-fifths rule); shadow testing, in which human reviewers perform the same task as the AI Tool and outcomes are compared; testing against representative or synthetic data sets; and proxy methods such as Bayesian Improved Surname Geocoding (BISG) where direct demographic data is not feasible.[22]

- **Navigating Practical Challenges for Self-Testing.** Developers and Deployers should be aware of, and build proactive solutions for, the following common practical challenges:
  - ***Consent for Sensitive Data.*** In many jurisdictions, the use of demographic data (such as race or gender) about individuals is considered "sensitive data" and requires consent for use, including testing. In such cases, testers can sometimes rely on privacy-enhancing techniques to enable the more secure and lawful use of demographic data where it is available, use synthetic data sets, or turn to less data-dependent methods, such as human shadow comparisons and red teaming.
  - ***Privilege.*** In some jurisdictions, assessments will be conducted under attorney-client privilege where available, in order to enable developers and deployers to feel confident engaging in this best practice.
  - ***Assessing for Disparate Impact.***[23] Determining whether a particular test result shows meaningful bias can be challenging. In some cases a test may be able to utilize disparate impact testing, in other cases, a tester may look to industry practices for relevant statistical deviations.[24] In addition, there will continue to be challenges identifying whether any bias identified by testing is caused by the test or by other conditions or factors; while this is most often relevant in the context of legal compliance, testers should consider the broader facts and circumstances with a view towards avoiding the test advancing bias in the tested population.

3. TRANSPARENCY

Provide Meaningful and Accurate Information to Affected Individuals

BASELINE PRINCIPLES

- Developers and Deployers each have important roles in ensuring that individuals subject to consequential employment decisions understand when, and to what extent, AI Systems are used to make consequential decisions, how those tools function, and what rights and options are available. (see Part II, Factor 4 – "Consequentiality").
- Disclosures should be provided by the entity that is best positioned to develop the content of the disclosure and communicate it to affected individuals.
- **Deployers** typically maintain a direct relationship with individuals and are the ones responsible for making or implementing consequential employment decisions. As a result, they primarily bear the responsibility for disclosures to individuals. **Developers**, meanwhile, are primarily responsible for providing Deployers with the information necessary to make those disclosures accurate and complete.
- *Obligations in this section are informed by the NIST AI Risk Management Framework and related guidance (including Section 3.4 on Transparency and Accountability), as well as ISO/IEC 42001, relevant provisions of the EU AI Act, and other relevant AI governance laws, regulations, and guidance.*

3.1 Developer Transparency

Developers are generally best positioned to provide information about how an AI System was designed, trained, and tested, its intended uses, any limitations of its effectiveness or restrictions on use, and how it may be configured or adapted. Their primary responsibility is to provide information to Deployers, in order to equip Deployers to deploy the AI System responsibly and to communicate about it accurately in the specific context of its deployment.

General Obligations of Developers

- **Inform Deployers:** Provide Deployers with specific disclosures (described below) that are sufficient to enable responsible deployment, and for Deployers to provide accurate, relevant, and compliant downstream disclosures to individuals. Disclose to Deployers when serious incidents or new material risks are discovered through post-deployment monitoring and testing. Where possible, developers should also consider rendering assistance by sharing information with deployers in the event of a regulator inquiry.[25]
- **Inform the public:** Publish accessible information about their responsible AI program and the general characteristics of their AI Systems, for example on a public website or in a high-risk AI System database where required by law.[26]
- **Substantiate claims:** Avoid representations that an AI System is "bias free," "objective," or otherwise immune from discriminatory outputs. Claims about tool accuracy, fairness, or efficacy should be grounded in documented evidence.

- **Use standardized formats:** Where feasible, use standardized documentation formats such as model cards and system cards to communicate technical information about AI Systems to Deployers.[27]
- **Facilitate transparency in the user interface:** Where the Developer controls the interface through which an Individual interacts with an AI System, the Developer should build in appropriate disclosures or disclosure functionality that the Deployer can enable.

EXAMPLE: FACILITATING TRANSPARENCY IN THE USER INTERFACE

A recruiting software developer offers an AI-powered interview platform that candidates access directly through the developer's hosted application. Because the developer controls the interface — not the employer — it builds a persistent on-screen notice stating "This interview is conducted by an AI system; your responses will be evaluated automatically and shared with the employer" with a link to a plain-language explanation of how the tool works. The deployer remains responsible for informing candidates, before directing them to the platform, that AI-assisted interviewing is how it will be used, and what rights they have with respect to that evaluation (e.g., a right to explanation; right to contest the decision; right to obtain a human intervention).

Specific Developer Disclosures to Deployers

Developers should specifically disclose to Deployers:

- **Intended purposes:** the purposes for which the tool is designed, and any uses that are restricted or prohibited;
- **Known limitations:** intended deployment conditions and any known limitations, including conditions under which performance may degrade, or populations or use cases for which the tool has not been validated;
- **Training:** a general description of how the tool was trained;
- **Testing performance:** the system-level safeguards, including testing procedures and performance metrics (e.g., likelihood of data leakage), to support informed deployment.[28]
- **Risk/Impact assessment:** any measures taken by the Developer to assess and/or mitigate potential risks arising from the use of the AI System, such as reasonably foreseeable risks of unlawful discrimination;
- **Data practices:** whether the tool uses Deployer or Individual data to generate outputs or for model training or improvement, and under what conditions; and
- **Configuration and oversight guidance:** instructions for appropriate deployment, including any Developer recommendations regarding human oversight during operation.

EXAMPLE: MODEL CARDS

A developer offering an AI-powered resume screening tool publishes a model card disclosing that the tool was trained on resumes from a particular industry sector and has not been validated for use with candidates from nontraditional educational backgrounds. A logistics company deploying the tool uses this information to configure additional human review for applications the tool flags, and to disclose accurately to applicants that AI screening is in use and subject to human review.

3.2 Deployer Transparency

Deployers maintain direct relationships with individuals (e.g. employees, applicants, candidates) and bear responsibility for consequential employment decisions. As a result, Deployers are generally best positioned to provide information to Individuals about the use of an AI System in shaping, informing, or making a consequential decision, and what rights Individuals have.

Deployer disclosures should be accessible, non-technical, and accurate — reflecting the Deployer's actual implementation of the tool, not merely the Developer's generic documentation. Where a Deployer configures or uses an AI System differently than the Developer intended, that departure should be disclosed to Individuals. Deployers who combine tools from multiple vendors, or who build workflows that incorporate AI at multiple stages, bear disclosure responsibility for the resulting process as a whole.

General Obligations of Deployers

- **Incorporate Developer information:** Deployer disclosures to Individuals should incorporate relevant information provided by Developers about how the AI System was designed and trained, to the extent that information bears on the Deployer's use and the Individual's situation.[29]
- **Describe the role of the tool:** Provide information about how the AI System operates, the categories of input data it uses, how its outputs factor into the decision making process, and what human review is applied.
- **Disclose departures from Developer guidance:** If the Deployer's configuration or use of the AI System differs from the Developer's instructions or documented intended use, that difference should be described in disclosures to Individuals.
- **Ensure disclosures are timely:** Individuals should receive notice before or at the point at which an AI System begins to affect their evaluation. For tools used in the initial application process, notice should be provided at the point of application or earlier. For tools used in interview or assessment stages, notice should be provided before that stage begins.[30]
- **Inform Individuals of their rights:** Disclose what rights Individuals have with respect to the use of the AI System (such as opt-outs, where applicable), how to exercise those rights, and how Individuals with disabilities may seek reasonable accommodations.

Specific Deployer Disclosures to Individuals

Deployers using AI Systems in consequential employment decisions should disclose to Individuals:

- **Fact of AI use:** that an AI System is being used, described in plain language;
- **Role of the tool:** how the AI System is used in the process — for example, whether it screens resumes, scores assessments, conducts initial interviews, or informs final decisions;
- **How outputs influence decisions:** how the tool's outputs are factored into the consequential decision, and the extent to which human review is applied;[31]
- **Data inputs:** the general categories of data the tool uses, and whether Individual data is used for model training or improvement;
- **Risk mitigation measures:** any steps taken to assess or reduce the risk of unlawful discrimination or other harms;
- **Rights and accommodations:** what rights Individuals have with respect to the AI System's use — including the right to request human review where applicable — and how Individuals with disabilities may seek accommodations; and
- **Departure from Developer guidance:** whether the Deployer's configuration or use of the AI System differs from the Developer's instructions or documented intended use, and how.

EXAMPLE

A financial services firm uses an Applicant Tracking System (ATS) that integrates three AI components from different vendors: a resume parser, a skills-matching engine, and an interview scheduling assistant. The firm's external job posting discloses that AI tools are used in the screening process and links to a plain-language summary of how each component is used and how applicants can request human review. The disclosure reflects the firm's actual multi-vendor workflow, not the documentation of any individual vendor.

3.3 Shared Principles for Effective Transparency

The following principles apply to disclosures by both Developers and Deployers.

Audience-Appropriateness

Both Developers and Deployers should ensure the information they provide is tailored to the appropriate audience and reflects a useful level of detail and technical complexity. For example, a Developer's disclosures to a Deployer may contain technically complex information that is relevant to Deployers but is generally not useful for individuals. Deployer disclosures to individuals, on the other hand, should use plain language calibrated to a general audience.

Standardized Formats

As the field of AI governance has matured, significant work has been dedicated to standardizing the format in which transparency information is provided, particularly between Developers and Deployers.[32] Standardized formats such as model cards and system cards provide a common language for entities to communicate about AI Systems and ensure transparency is consistent across contexts and geographies. Developers should use standardized formats where feasible.

Confidentiality Limits

Transparency obligations do not require disclosure of trade secrets, proprietary model weights, or implementation details that could enable security vulnerabilities or adversarial manipulation of AI Systems.[33] Where confidentiality concerns limit the depth of disclosure on a particular point, organizations should err toward providing more general information rather than no information.

3.4 Generative AI and Agentic Systems: Further Considerations

Transparency obligations take on particular characteristics when AI Systems generate natural-language outputs, such as written evaluation summaries or recommendations, or when agentic AI Systems execute multiple steps across the hiring workflow.

- **AI-generated evaluation content:** Deployers should ensure that AI-generated evaluation content — candidate summaries, assessment narratives, interview feedback — is clearly identified as such to human reviewers. Human reviewers should understand that they are exercising independent judgment on AI-generated material.
- **Agentic AI pipelines:** When an AI System executes multiple steps autonomously across the hiring workflow — for example, sourcing candidates, screening resumes, scheduling interviews, and ranking applicants — the Deployer's disclosure obligation extends to the overall workflow. Deployers operating agentic pipelines should be able to describe, at a workflow level, how decisions are allocated across automated and human steps, and should provide that description to individuals as part of pre-use disclosure.
- **Non-deterministic outputs:** Probabilistic models that underpin Generative AI and Agentic Systems may produce materially different outputs for similarly situated candidates. Developers should disclose this characteristic to Deployers. Deployers should account for it when describing how the tool operates, and should not represent such outputs as consistent, objective, or bias-free.[34]
- **Explainability:** Transparency obligations are complicated by a fundamental limitation of large language models and many other AI Systems: they cannot always produce a traceable, reliable account of why a specific output was generated for a specific individual. Technical, academic, and regulatory communities[35] are actively grappling with this tension. Where full model-level explanation is not technically feasible, Deployers should be able to describe the categories of inputs that influenced the output, the role that output played in the decision, and what factors a human reviewer considered. Developers bear responsibility for designing systems that support this capacity, and for disclosing to Deployers where and why that capacity may be limited.

EXAMPLE: AGENTIC SOURCING, SCORING, AND OUTREACH

An employer uses an agentic recruiting assistant that automatically sources candidates from professional networks, scores resumes against a job description, generates draft outreach emails, and ranks candidates before any human review occurs. The employer's careers page and job postings provide layered disclosures that AI tools are used throughout the recruiting process, including to identify and rank candidates, and when human review is applied. The disclosure describes the workflow as a whole, in sufficient detail for applicants to understand, and the role of human review.

4. DATA SECURITY AND PRIVACY

Implement Safeguards to Protect Personal Data and Maintain the Integrity of AI Reasoning and Agency

BASELINE PRINCIPLES

- The sensitive nature of employment decisions obligates developers and deployers that process personal data to protect the confidentiality and integrity of personal data while at rest and in transit.
- Developers and Deployers should institute systems and structures around AI Systems to prevent personal data from being used in a way inconsistent with individuals' choices, expectations, legal obligations, and established privacy policies.
- In addition to comprehensive privacy and security laws, AI Systems must also comply with specific employment related legal obligations, such as the Fair Credit Reporting Act (FCRA), when AI Systems are utilized for U.S. background checks or other workplace assessments that fall under that jurisdiction.
- *Obligations in this section are informed by the NIST AI Risk Management Framework and related guidance, the NIST Cybersecurity Framework (CSF) 2.0, the NIST Privacy Framework 1.0, ISO/IEC 42001, relevant provisions of the EU AI Act, and other relevant privacy, cybersecurity, employment, and AI governance laws, regulations, and guidance.*

4.1 Comprehensive Privacy and Security Programs

Developers and Deployers of AI Systems, including Generative AI and Agentic AI Systems, should maintain comprehensive privacy and security programs designed to protect personal data. These programs must include administrative, technical, and physical safeguards that are appropriate to the sensitivity of the information handled.

Beyond standard protections, these programs should address risks arising from the use of large language models, Generative AI and Agentic AI, when applicable. For example, this may require specific defenses against prompt injection to prevent malicious inputs from bypassing model filters or manipulating outputs. Programs must also protect against data leakage and model inversion to prevent training data or sensitive user inputs from being extracted or disclosed to unauthorized parties. Because Agentic AI draws strength from its ability to act on data across different parts of the business — including internal platforms and third-party tools — security programs must evaluate security risks from the linked systems and the data flows involved.

4.2 Specific Security and Accountability Measures

To protect the integrity of the data environment and enforce accountability, organizations should apply precise identity permissioning, comprehensive observability, and data security measures across data sources. Effective security practices include:

- Secure storage of personal data to prevent unauthorized access or accidental loss.
- Encryption of personal data when in transit.
- Identity permissioning and strict access segmentation for sensitive personal data to prevent exposure to unauthorized personnel.
- Comprehensive logging and monitoring of system usage and agentic data flows to create an audit trail for security incidents.
- Clear contractual obligations to define security duties between Developers and Deployers.
- Guarding against injection attacks and adversarial risks that target the model architecture or its linked systems.
- Regular privacy and security training and accountability measures to reduce the risk of human error or social engineering.
- With respect to agentic systems, establishing technical measures to verify and authenticate agent identities and data across the entire supply network.

4.3 Training on Personal Data

The use of personal data for model development and refinement must be designed to prevent inappropriate use.

- Training or fine-tuning AI Systems on personal data should only be conducted when there is an approved lawful basis, such as consent, legitimate interest or a legal provision, and for specified, legitimate purposes. Where practicable, training should use data that has been de-identified or subject to PETs in order to minimize the processing of personal data.
- Developers should establish clear standards for what data should or should not be included in these sets, taking into account risks such as proxy bias and the need for diverse, representative data.
- In the context of the AI value chain, this framework must recognize the relationship between controllers and processors. When a Developer acts as a processor, it should only perform training or fine-tuning in accordance with the controller's documented instructions and applicable law.
- To support transparency, the use of personal data to train or improve an AI System must be documented and conducted in a manner that prevents unreasonable risks of data disclosures. Developers should also give Deployers the means to understand how personal data obtained from them is being utilized.

4.4 Privacy and Sensitive Data Inference

Organizations must be vigilant regarding the automated reasoning capabilities of AI, which can create new privacy risks. Because AI can process vast amounts of data, there is a risk that tools may be used to infer sensitive data that was never explicitly provided by the individual. Developers and Deployers should implement controls to guard against this risk and to ensure systems are developed and deployed consistent with the principle of data minimization.

5. HUMAN OVERSIGHT AND AGENTIC AUTHORITY LEVELS

Preserve Meaningful Human Oversight and Calibrate Agentic Authority to Risk

BASELINE PRINCIPLES

- Human oversight of an AI System is not satisfied by simply being "in-the-loop." Instead, AI Systems should be designed and deployed in ways that ensure humans have the *information, authority, capability, and genuine opportunity* to understand, question, approve, override, or halt AI-influenced actions and outcomes.
- Developers should design AI Systems in ways that allow Deployers to implement meaningful oversight mechanisms. Deployers should determine and implement the appropriate governance model for the specific situation, workflow, and use case.
- *Obligations in this section are informed by the NIST AI Risk Management Framework and related NIST guidance on agentic AI, as well as ISO/IEC 42001:2023, relevant provisions of the EU AI Act, and other relevant AI governance, employment, and emerging AI laws, regulations, and guidance on human oversight and autonomous AI Systems.*

5.1 Human Oversight

AI Systems should be designed and deployed in ways that ensure humans have the *information, authority, capability, and genuine opportunity* to understand, question, approve, override, or halt AI-influenced actions and outcomes. Effective AI oversight models should take into account whether the system is deterministic, generative, or agentic, and the totality of the risk profile identified in Part II (i.e., the sensitivity of data and inferences, degree of independent judgment, proximity to the decision, and nature and severity of impact).

Designing oversight mechanisms with these factors in mind ensures AI Systems augment human judgment, and do not obscure the human's responsibility.

5.2 AI Authority Levels

Effective AI oversight models must balance the benefits of AI adoption with the risks inherent with AI use. As a result, Developers and Deployers should consider governance approaches which recognize that AI may be granted varying levels of authority based on the risks involved.[36] These frameworks should define the appropriate level of authority for a given AI interaction, and reflect a holistic understanding of the system, including how material the AI interaction is to the ultimate consequential decision. Those varying levels may include:

1. **Fully autonomous agentic authority not allowed.** A human determines all aspects of the workflow, including whether or how AI may assist, whether action is warranted, and how to execute that action, because the matter is too risky, too consequential, insufficiently governable, legally restricted, or otherwise unsuitable for AI execution.

2. **AI advises and humans decide.** AI may generate content, scores, classifications, or recommendations, but a human retains full decision-making authority and must manually execute any consequential action.
3. **AI acts with human approval.** AI may prepare or recommend an action, but the action may not be executed unless a human affirmatively approves it in advance.
4. **AI acts with human review.** AI may execute a bounded action, subject to human review after execution.
5. **AI acts fully autonomously.** AI may execute within defined boundaries, supported by continuous monitoring and automated secondary checks that surface outcomes for human review.

Which oversight model to choose depends highly on the context in which the AI is being used. For example, allowing AI to act without human review may only be appropriate for actions that are low- or moderate-risk, reversible, auditable, and subject to prompt escalation and remediation. Similarly, AI should only be allowed to be autonomous where the action is sufficiently low-risk or reversible, human accountability remains clear, and applicable law does not require direct human involvement, review, or recourse.

AUTHORITY LEVEL	EXAMPLE
Agentic Authority not allowed	Automated termination decision
AI advises	Creating rankings for hiring decisions (resume scores, candidate ranking)
AI with human approval	Updating an address, processing payroll
AI with human review	Schedule changes, time off request approval
AI autonomously	User alerts, report generation

Figure 2. Examples of how an organization may choose to define agentic authority levels.

AI Systems may involve several actions with different authority levels, or a structure where one AI has control over other AI Systems. When multiple authority levels exist within a single system, the system should be reviewed in its entirety to ensure that actions taken with a lower level of authority are not bypassing or overriding higher authority level requirements elsewhere.

The decision to grant any particular authority level to an AI System should not be set in stone. All AI Systems, and especially those that have heightened risk profiles, should be tested, monitored, and revalidated periodically to ensure that the AI System's authority level, controls, and safeguards remain appropriate.

5.3 Meaningful Human Accountability

Regardless of how much authority is delegated to AI Systems and tools, humans must always remain accountable for the resulting action. At a minimum, then, meaningful human oversight requires:

- clear assignment of responsibility and decision rights;
- training sufficient to recognize limitations, errors, and automation bias;
- documented escalation, override, and appeal mechanisms;
- logs and records sufficient to support auditability, traceability, and review; and
- periodic reassessment of whether the selected oversight model remains appropriate as the system, use case, or risk profile changes.

Practical considerations should also be taken into account. For example, meaningful oversight would require preventing otherwise reversible actions becoming irreversible due to a delay between the AI System's output and the human's review or approval.

5.4 Developer and Deployer Responsibilities

Developers and Deployers must work in concert to effectively govern AI, and have distinct responsibilities in relation to designing and deploying AI Systems.

Developers should design AI Systems in ways that allow Deployers to implement meaningful human oversight mechanisms. Key oversight mechanisms include:

- Providing system functionality that enables deployers to exercise meaningful human oversight, for instance through configurable levels of authority and access controls;
- documenting intended use, limitations, and known risks;
- supporting transparency and explainability, or providing compensating controls where those are limited;
- providing intervention, rollback, and override mechanisms; and
- enabling logging, monitoring, and auditability.

Deployers should determine and implement the appropriate governance model for the specific situation, workflow, and use case. This includes:

- validating that the selected authority level is lawful and appropriate;
- monitoring outcomes for unfairness, error, and drift;
- preserving required notices, accommodations, appeals, and human review; and
- ensuring that a human decision-maker remains accountable for higher risk or consequential employment outcomes.

ENDNOTES

1 Future of Privacy Forum, *Best Practices for AI and Workplace Assessment Technologies* (2023). The 2023 Best Practices focused on the use of predictive AI systems in consequential employment decisions, including recruiting, hiring, promotion, and termination.

2 NIST, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1 (July 2024), https://doi.org/10.6028/NIST.AI.600-1. The Profile identifies twelve risk categories unique to, or exacerbated by, generative AI — including confabulation (hallucination), information integrity, information security, and data privacy.

3 See e.g., the Colorado AI Act, Illinois H.B. 3773, the EU AI Act (Arts. 5(1)(f) and 10(5)), New York City Local Law 144, California FEHA Automated Decision System Regulations, California Privacy Protection Agency Automated Decisionmaking Technology Regulations, and GDPR Art. 22.

4 NIST AI RMF, NIST AI 600-1 (Generative AI Profile), ISO/IEC 42001, and ISO/IEC 42005.

5 This example is discussed in the European Commission's *Draft Guidelines on the Classification of High-Risk AI Systems under the AI Act* (May 19, 2026). Sections 3.4 and 3.4.2(v)(b) suggest that Annex III does not apply to AI systems that recommend job opportunities or analyze candidate-provided information.

6 European Union, *Regulation (EU) 2024/1689 (EU Artificial Intelligence Act)* (2024). The AI Act classifies many AI systems as high-risk based on their "intended purpose" in the contexts identified in Article 6 and Annex III, including employment-related use cases.

7 It has been demonstrated that LLMs are highly capable of inferring sensitive personal attributes from unstructured text even when those attributes are not directly stated. Robin Staab, Mark Vero, Mislav Balunović & Martin Vechev, *Beyond Memorization: Violating Privacy Via Inference with Large Language Models* (revised May 2024).

8 In multi-step agentic workflows, individual errors or small rates of inaccuracy can compound ("compounding errors"), while gains in single-step accuracy can compound into exponential improvements. See, e.g., Anthropic, *Building Effective Agents* (Dec. 19, 2024). For a more technical source, see Sinha, Arun, Goel et al. (2025), *The Illusion of Diminishing Returns: Measuring Long Horizon Execution in LLMs,* arXiv:2509.09677 (Cambridge / Stuttgart / MPI).

9 It should be noted that an output that is functionally determinative — for example, an early-stage screen that removes most applicants from further consideration — can drive the outcome for those it affects, even though it may sit far from a final offer of employment. See Manish Raghavan & Solon Barocas, *Challenges for Mitigating Bias in Algorithmic Hiring*, Brookings Institution (Dec. 6, 2019) (explaining that algorithmic screening can be highly consequential because it operates as a major filter through which applicants must pass).

10 For Developers, this may include both customer-facing AI Systems as well as their own AI Systems built or procured for internal use.

11 This definition mirrors NIST AI 600-1, the Generative AI Profile of the AI RMF (July 2024).

12 There is no consensus legal definition of "agentic AI." This definition draws from the OECD, *The Agentic AI Landscape and Its Conceptual Foundations*, (OECD Artificial Intelligence Papers No. 56, 2026), which emphasizes autonomous, goal-directed behavior, multi-step planning, and the capacity to execute action sequences with limited human intervention. In the United States, NIST/CAISI, in its *January 2026 Request for Information*, characterizes AI agent systems as capable of taking autonomous actions that impact real-world systems or environments.

13 These four functions track the NIST AI Risk Management Framework (AI RMF 1.0), NIST AI 100-1 (Jan. 2023); NIST frames Govern as a cross-cutting function intended to be infused throughout Map, Measure, and Manage rather than performed sequentially.

14 The NIST National Cybersecurity Center of Excellence (NCCoE), *Accelerating the Adoption of Software and AI Agent Identity and Authorization* (Feb. 5, 2026). The concept paper highlights the need for identity, authorization, auditing, and prompt injection controls for AI agents operating across enterprise systems and interacting with enterprise data, tools, and applications.

15 Impact assessments are a core accountability tool across emerging AI governance frameworks. See ISO/IEC 42005:2025 (AI system impact assessment), which provides guidance on conducting and documenting AI system impact assessments — including foreseeable misuse — across the lifecycle and is designed to complement ISO/IEC 42001; and the Map function of the NIST AI RMF. For a survey of impact-assessment requirements, see Future of Privacy Forum, *AI Governance Behind the Scenes: Emerging Practices for AI Impact Assessments* (2025 Update).

16 On generative-AI-specific risks, see the NIST Generative AI Profile, NIST AI 600-1 (July 2024), which identifies twelve risk categories — including confabulation/hallucination, data leakage, prompt injection, and information-integrity harms — each mapped back to the AI RMF's four functions.

17 Drift refers to gradual degradation of a model's predictive accuracy over time as real-world data diverges from the data it was trained on. See IBM, https://www.ibm.com/think/topics/model-drift.

18 Escalation is often contextual and proportionate to an organization's size and resources. In contrast to feedback channels (such as an anonymous ethics hotline), escalation should create a pathway to those who have the ability to mitigate or accept the risk - IT teams, leadership, governance committee, or other appropriate stakeholders.

19 On differentiated and adversarial testing (including red-teaming) for generative and agentic systems, see the NIST Generative AI Profile (NIST AI 600-1) and the Measure function of the NIST AI RMF.

20 On AI bias, see NIST, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, NIST SP 1270 (Mar. 2022), which frames AI bias as a socio-technical risk involving systemic, statistical / computational, and human sources of bias, and notes that harmful bias may arise regardless of intent because zero bias risk is not achievable.

21 For further reading, see, e.g., Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. Legal Analysis 113 (2018).

22 Consumer Financial Protection Bureau (CFPB), *Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity* (Sept. 2014). The report describes Bayesian Improved Surname Geocoding (BISG) and related proxy methodologies that may be used to estimate demographic characteristics when direct demographic data is unavailable. *See also* Center for Democracy & Technology, *Navigating Demographic Measurement for Fairness and Equity* (Apr. 2024). The report surveys contemporary techniques for demographic measurement and proxy-based approaches used to evaluate AI systems for bias and fairness.

23 The legal landscape governing disparate impact analysis in employment is evolving and varies across jurisdictions. In the United States, Executive Order 14281 (Apr. 23, 2025) directs federal agencies to deprioritize enforcement, and a June 9, 2026 DOJ Office of Legal Counsel opinion found the EEOC's Title VII guidelines constitutionally suspect. These positions do not directly amend Title VII, under which disparate impact liability remains codified and available to private plaintiffs and many state enforcement authorities, or affect requirements that may arise from state civil rights laws and employment-specific regulations, such as New York City Local Law 144 (2021).

24 An example of one approach to assessing disparate impact in testing is from New York City's law addressing bias in AI-driven employment tools. NYC Local Law 144 (2021).

25 For example, EU AI Act Article 13(1)-(3) requires provider-to-deployer instructions and information for high-risk AI systems, provider cooperation with competent authorities, and deployer cooperation with competent authorities.

26 For example, EU AI Act Articles 49 and 71 require certain high-risk AI systems listed in Annex III, including those used in employment contexts, to be registered in a public EU high-risk database. Similarly, NYC Local Law 144 requires public posting of AEDT bias-audit summaries. For frontier models, see the New York RAISE Act, which requires covered frontier-model developers to publish safety-protocol information, subject to appropriate redactions.

27 These formats provide a common structure for disclosing design decisions, training data, known limitations, and intended use, and their adoption is increasingly expected under frameworks including the NIST AI RMF (Govern 1.1, Map 1.1) and ISO/IEC 42001.

28 See Singapore Personal Data Protection Commission (PDPC), *Proposed Advisory Guidelines on the Use of Personal Data in Generative AI* (June 2026). Paragraph 8.3 recommends that system providers share information with downstream deployers regarding system-level safeguards, including testing procedures and performance metrics (e.g., the likelihood of data leakage), to support informed deployment and the protection of personal data.

29 Although EU AI Act Article 13 applies to providers rather than deployers, it requires providers to furnish instructions for use describing an AI system's characteristics, capabilities, limitations, intended purpose, and conditions of appropriate use. Together with NIST AI RMF 1.0 (GOVERN 1.1, MAP 1.1) and ISO/IEC 42001, this documentation provides a foundation for deployers to incorporate relevant developer information into disclosures provided to affected individuals.

30 The Colorado AI Act, SB24-205, §§ 6-1-1703 to 1704 requires deployers of high-risk AI systems to provide notice to individuals before or at the time the system makes, or is a substantial factor in making, a consequential decision.

31 The European Commission's Draft Guidelines on High-Risk AI Systems, Annex III, Point 4(a)-(b) recognize that AI systems may materially influence employment decisions even where humans formally make the final determination if those decision-makers significantly rely on AI outputs.

32 NIST's AI Standards "Zero Drafts" Pilot Project reflects ongoing efforts to develop standardized guidance for model and dataset documentation to promote transparency among AI actors.

33 The Colorado AI Act provides that transparency obligations do not require disclosure of trade secrets or other information protected from disclosure under state or federal law, provided the consumer is notified when information is withheld. See Colo. Rev. Stat. § 6-1-1704(5). See also Connecticut Pub. Act No. 26-15, § 8.

34 The FTC's final order in *In the Matter of IntelliVision Technologies Corp*. prohibits unsupported claims that AI systems are accurate, effective, or bias-free.

35 The EU AI Act, California ADMT Regulations, and Colorado AI Act each emphasize meaningful transparency, disclosure of the role and operation of AI systems, and human oversight without requiring reconstruction of the precise internal reasoning underlying a particular AI-generated output.

36 The NIST National Cybersecurity Center of Excellence (NCCoE), *Accelerating the Adoption of Software and AI Agent Identity and Authorization* (Feb. 2026), concept paper explores how organizations can govern AI agents with varying degrees of autonomy through identity, authorization, delegation, and emphasizes that governance controls should be calibrated to the actions an AI agent is authorized to perform independently.

1350 EYE STREET NW, SUITE 350 | WASHINGTON, DC 20005 FPF.ORG/AI | info@fpf.org