

OECD AI Transparency Report

Organization: Salesforce (USA)

Reporting Period: Q3 2026

Published: October 2, 2026

Section 0 – Your role in the AI system lifecycle

Responding as (please select all that apply):

- Application developer (/provider)
- Model developer (/provider)
- Deployer

If any of the fields above are selected, please indicate your organization size:

Large enterprise: 250 or more employees

Section 1 – Risk Identification and Evaluation

1. How does your organization define and/or classify different types of risks related to AI?¹

- Maintain an internal taxonomy, classification system, and/or set of risk area definitions
- Align with specific external frameworks, standards, reports, and/or guidance documents such as taxonomies, classification systems, and/or sets of definitions for risk areas (e.g. the NIST AI Risk Management Framework, the ISO/IEC 42001, the OECD AI Principles, the EU General-Purpose AI Code of Practice, the International AI Safety Report). (please specify): Salesforce aligns its risk governance with the NIST AI Risk Management Framework, and applicable legislation and the UN Guiding Principles on Business and Human Rights. We also achieved ISO 42001 certification (the international AI-management standard) for Agentforce, Einstein, and Slack AI.
- Consult with external stakeholders (e.g., from academia, industry groups, civil society, or standard development organisations) when developing or updating risk taxonomies

How does your organization define thresholds at which severe risks, unless adequately mitigated, may be considered "unreasonable"? Please describe the framework(s) most relevant to your organization's role and context. If your organisation does not use formal thresholds, you may describe instead the policies, decision criteria, or governance processes you apply when assessing severe AI risks. You may also describe:

- Specific examples of contexts in which your models or systems could, in your assessment, pose unreasonable or intolerable risk;
- How mitigations are graduated by capability or propensity level;
- How your organisation considers the specificities of different AI technologies, such as generative AI and agentic AI, including any specific risk assessment or governance mechanisms applied to agentic AI capabilities.

Salesforce scores risks using an automated due diligence framework that classifies products as either lower risk or higher risk. Mitigations are graduated based on these thresholds:

- **Lower-Risk Products:** Teams are fast-tracked and provided with detailed self-service enablement to develop AI in line with legal and ethical requirements
- **Higher-Risk Products:** These require deep human review. The Responsible AI and Technology team initiates kick-off meetings, conducts thorough risk assessments, evaluates the impact with and without mitigations, and requires re-testing post-mitigation.

Salesforce utilizes an automated due diligence framework that classifies AI products as lower or higher risk. This process is led by Responsible AI and Tech (RAIT) product managers within the Office of Ethical and Humane Use (OEHU).

- **Threshold Definitions:** Risks are identified and categorized under specific sociotechnical harm subtypes. We define unreasonable risk as any unmitigated output that results in catastrophic damage, individualized professional advice (medical/legal/financial) without human review, or the prediction of protected characteristics.
- **ExploitAI Framework:** Our Security team performs ExploitAI activities to uncover AI-specific vulnerabilities, including hallucination exploitation, knowledge poisoning, and multi-agent exploitation.
- **Empirical Success:** Through these pre-launch assessments, we have successfully reduced toxic or unsafe outputs by 35% in specific AI features and decreased biased outputs in partner-integrated content from 69% to 4%.

2. Both before and after deployment, what practices does your organization use to identify and evaluate vulnerabilities, incidents, emerging risks and misuse throughout the AI system lifecycle?

- Internal red teaming³
- Qualitative assessments (e.g., expert judgment, narrative scoring)
- Internal benchmarks, standard metrics (e.g., frequency, severity scores)
- Impact assessments such as personal data protection assessment and fundamental rights impact assessments
- Automated monitoring / remote status reporting (e.g., telemetry pipelines, automated alerts when risk thresholds are breached)
- Specific risk identification practices for agentic AI capabilities, e.g., autonomous action-taking, tool use, multi-step planning (please specify): For autonomous agents, Salesforce evaluates responses using specific metrics like Instruction Adherence (to check if instructions were followed) and Task Resolution (to confirm the user's overall task was addressed).
- Multilingual and cross-cultural evaluation, including collaboration with local ecosystems (e.g., national AI institutes, language-specific evaluation consortia) for under-represented languages and cultural contexts (please specify): Voice guardrails are tested to ensure Automatic Speech Recognition and Text-to-Speech models operate accurately across all languages, dialects, accents, and grammars.

Please briefly describe your organization's decision-making process for deploying an AI model or system including: who has decision authority; what evidence and criteria are required (e.g., performance, safety/security, robustness, fairness, privacy, compliance, independent external testing,

evaluation, validation and verification (TEVV) or assurance); contextual caveats or limitations required in interpreting risk evaluation metric results; what approvals/sign-offs are needed; and what safeguards (e.g., mitigations, monitoring, or phased release) must be in place to move into deployment.

Salesforce mandates that all internal and customer-facing AI products undergo ethics reviews before they can be released to the market. The decision-making process follows a standardized flow:

- **Intake & Triage:** Product teams submit targeted intake forms, which are automatically scored to designate the product as lower or higher risk.
- **Review & Testing:** For higher-risk products, Responsible AI product managers initiate a deep evaluation, categorizing risk scenarios both with and without mitigations. The Testing, Evaluation, and Alignment team conducts tests using both qualitative expert judgment and quantitative industry thresholds.
- **Implementation & Re-testing:** Product teams must implement identified technical and non-technical mitigations, and the product must be re-tested post-mitigation before deployment.

Governance and Approvals Oversight and sign-off authority are distributed across dedicated governance bodies. The Salesforce internal AI Trust Council, composed of senior executives from engineering, legal, security, and ethics, accelerates decision-making for AI products. Internal AI tools require sign-off from the Internal AI Governance Review team. Ultimately, the Chief Ethical and Humane Use Officer reports directly to the C-Suite, and the Board of Directors' Cybersecurity and Privacy Committee is primarily responsible for reviewing AI strategies, policies, and priorities.

Following deployment, how do you use incident reports — including those shared by other organizations — to identify emerging issues, update risk assessments, and trigger corrective actions.

Following deployment, Salesforce uses a dedicated Ethical Use Policy team to manage case operations, content safety, and user escalations. Workspace users in regions like the EU can use in-product reporting tools (e.g., Report to Slack) to flag prohibited or illegal content, which is then reviewed around the clock by specialists.

The Policy team uses these reports and stakeholder engagement to actively identify emerging risks, potential harms, or policy gaps. These insights trigger a rigorous update process for customer-facing guidelines; for example, the AI Acceptable Use Policy (AUP) is continuously refined as risks and technologies evolve. Additionally, to trigger rapid corrective action for broader security issues, Salesforce employs an AI-powered incident response system that automatically triages next steps based on case descriptions, saving thousands of hours and evaluating incident response quality.

3. Does your organization have mechanisms to receive and/or share reports of risks, incidents or vulnerabilities by/with third parties?

Yes, both receiving and sharing

If yes, please indicate which stakeholders reporting mechanisms may be accessible to, recognising specifics will depend on the nature of the report and risk level. For each, indicate whether access is on a standing basis (Always), activated based on context (Ad hoc/as relevant), or not currently accessible.

	Always	Ad hoc / as relevant	Not currently accessible
Relevant employees (security, legal/compliance, etc.)	☐		
Public authorities (e.g., public reporting offices for security incidents or data breaches)		☐	
Supervisory authorities (e.g., data protection authorities)			
Industry partners or industry fora		☐	
Users (end users, enterprise customers, developers using APIs)	☐		
Business customers / deploying organizations (if different from “users”)			
Supply-chain partners (cloud, data providers, etc.)			
Academia / researchers			
Independent auditors			
Standards bodies / multi-stakeholder initiatives (OECD.AI, ISO/IEC, IEEE, etc.)			
Specific civil society organizations		☐	
Media / general public (people or communities indirectly impacted)			
Other			

Please describe how your organisation determines when and how reporting mechanisms are made accessible to different stakeholders (e.g. based on legal or regulatory requirements, internal policies, or risk level). Please also provide public URLs or details on how to submit reports. If none, write “Not applicable”

Salesforce operationalizes mechanisms to receive reports of risk and misuse directly from our users. To comply with the European Union’s Digital Services Act (DSA), we introduced in-product reporting capabilities, including the Report to Slack option, which allows workspace users in the EU to seamlessly flag prohibited or illegal content. The Ethical Use Policy team and its global Operations specialists actively manage these user escalations, working around the clock to assess reports against safety standards and the company’s AUPs.

How does your organization share relevant information as appropriate?

- Information-sharing partnerships (with confidentiality agreements as appropriate)
- Reporting to regulatory authorities as relevant

- Other (specify): (Publishing to public web pages like trust.salesforce.com and the Salesforce Compliance website).

Please describe what type of information is shared.

Salesforce emphasizes transparency by sharing a variety of risk, vulnerability, and compliance data with external third parties:

- **Security Vulnerabilities and Advisories:** Security-related information, threat landscapes, and system advisories are shared publicly on the company's dedicated Security advisories page and trust.salesforce.com.
- **Model Limitations and Ethical Evaluations:** To help customers assess risk, Salesforce publishes structured AI Model Cards for its predictive, generative, and agentic AI models on its Compliance website. These documents openly share training data contexts, performance metrics, known ethical risks (like bias and toxicity), and use-case limitations.
- **Incident and Enforcement Metrics:** For platform safety transparency, Salesforce publishes detailed overviews of its DSA compliance efforts, including key metrics and enforcement actions taken against violative content, on its Compliance Resource Page.
- **Industry Safety Practices:** Salesforce actively shares emerging risk mitigation strategies with civil society and industry peers through best-practice blog posts, Trailhead training modules, and interactive digital resources like [A Day Empowered by Agents](#), which visually details how to orchestrate safe human-AI handoffs.

4. If your organization has programs to incentivize risk(s) disclosure, incidents and systemic vulnerabilities, please share what type of programs you use?

- Paid rewards ("bug bounty")
- Vulnerability Disclosure Programs (VDPs) with safe harbor / liability protection
- Research grants, contracts, and hiring red teamers
- Operational incentives for customers and partners
- Internal incentives and culture

Please provide public URLs and/or details on your incentive program(s). If none, write "Not applicable."

Salesforce does have formal programs to incentivize the disclosure of AI risks, vulnerabilities, and incidents. These efforts are primarily centralized through our Bug Bounty Program and our Office of Ethical and Humane Use, which have evolved to specifically address the unique risks of generative and agentic AI.

1. The Bug Bounty Program (Financial Incentives) - [Bug Bounty Program](#)

Salesforce operates one of the most active bug bounty programs in the tech industry, which explicitly includes AI-specific testing.

- **Expansion to AI:** The program incentivizes ethical hackers to identify not just traditional security flaws, but also AI bias, prompt injection, and harmful model behaviors.
- **Incentives:** Salesforce has paid out over \$23 million to date through its bug bounty network. In 2023 alone, they paid over \$3 million to roughly 650 researchers.

- **Focus on Agentic AI:** With the rise of autonomous agents (like Agentforce), the program now specifically targets systemic vulnerabilities where AI might autonomously make incorrect or insecure decisions without human input.

2. Responsible Disclosure & Ethical Reporting

Beyond financial rewards for hackers, Salesforce has structured frameworks for reporting and transparency:

- **HAIP Reporting Framework:** Salesforce participates in the OECD's Reporting Framework for AI System Developers. This involves documenting and disclosing how they evaluate models against trust and safety metrics such as privacy, robustness, and bias.
- **Adversarial Testing:** When AI models score poorly on safety metrics, Salesforce incentivizes internal and external teams to use red teaming (adversarial testing) to find and report how those risks manifest in practice.
- **Model Cards:** The use of Model Cards for our predictive and generative AI to disclose known risks, ethical considerations, and appropriate use cases to the public and our customers.

3. The Einstein Trust Layer

While not a reporting program in the traditional sense, this architecture acts as a systemic catch-all for vulnerabilities:

- **Feedback Loops:** The Trust Layer includes mechanisms for users to report toxic or incorrect AI outputs, which are then used to refine the models.
- **Data Masking & Audit Trails:** It provides a systemic disclosure mechanism by creating an audit trail of all AI interactions, allowing organizations to monitor and report incidents of data exposure or policy violations internally.

4. Vulnerability Disclosure Policy (VDP) - [Vulnerability Reporting](#)

Salesforce maintains a Vulnerability Disclosure Policy that provides a safe harbor for researchers. This program ensures that individuals who discover and report systemic AI vulnerabilities in good faith are not met with legal action, provided they follow the company's disclosure guidelines.

5. Does your organization use or contribute to the development of international technical standards, best practices or tools for the identification, assessment, and evaluation of risks?

Contributes and uses

If contributes or uses, please select all that apply:

- ISO/IEC standards (e.g., ISO/IEC 27001, ISO/IEC 42001, ISO/IEC 23894, ISO/IEC 42005)
- NIST AI Risk Management Framework (AI RMF)
- Industry-led best practices or codes of conduct
- Other international or regional standards (please specify): ISO 9001

6. Does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks that could have broad and potentially cascading socio-economic impacts? Please indicate which stakeholders you collaborate with, and whether

collaboration is on a standing basis (Always-on), activated based on context (Ad hoc/as relevant), or there is no current collaboration.

	Always-on	Ad hoc / as relevant	No current collaboration
Relevant employees (security, legal/compliance, etc.)	☐		
Public authorities / regulators / AI safety institutes / other public sector bodies and inter-governmental coordination structures		☐	
Industry partners or industry fora		☐	
Users (end users, enterprise customers, developers using APIs)			
Business customers / deploying organizations (if different from “users”)	☐		
Supply-chain partners (cloud, data providers, etc.)		☐	
Academia / researchers		☐	
Independent auditors		☐	
Standards bodies / multi-stakeholder initiatives (OECD.AI; ISO/IEC; IEEE; AIID etc.)			
Independent civil society organizations (e.g., for evaluations, audits, or policy review)		☐	
Media / general public (people or communities indirectly impacted)			
Other(s) (please add any other relevant stakeholder groups):			
Not applicable			

Please describe examples of how your organization collaborates. Examples may include: participating in multi-stakeholder or industry initiatives; partnering with academia, civil society, or the public sector; establishing internal or joint governance structures with external experts or in partnership with other efforts to share knowledge and jointly address risks.

Salesforce employs a multifaceted collaboration strategy to address AI risks, integrating internal oversight with global public-private partnerships. Internally, we maintain a cross-functional governance structure through our AI Council—uniting Legal, Security, and Ethical Use experts, to manage a centralized registry of internal agents, effectively balancing rigorous safety reviews with operational efficiency. In the regulatory sphere, Salesforce actively contributes to the development of international standards, participating in the EU AI Office’s Code of Practice for General-Purpose AI and advocating for interoperable, risk-based norms through direct consultation with government bodies. Our academic and civil society partnerships, such as our symposia with Stanford University, facilitate deep technical

exchanges on responsible agentic AI. Furthermore, our regional engagement means we work alongside governments and industry in Southeast Asia on national AI roadmaps and legislative briefings, while simultaneously convening multi-stakeholder calls to action in Latin America to address AI readiness and workforce development. Together, these initiatives represent a sustained, collaborative effort to harmonize industry innovation with the public interest and sovereign governance goals.

Any further comments and implementation documentation.

No answer provided

Section 2 – Risk Management and Information Security

7. Across the AI system lifecycle, what steps does your organization take to manage risks and vulnerabilities? Please indicate which specific practices, tools or metrics your organization uses at each lifecycle stage. You may refer to tools or metrics present in the OECD.AI Catalogue of Tools and Metrics or suggest additions. Respondents may complete only the lifecycle stages relevant to their organisation's role. Stages that do not apply may be left blank.

Plan and design

Collect and process data

Build, integrate and interpret model

Verify and validate

Deploy

Operate and monitor

Please provide a consolidated description of your organisation's approach to risk management across the AI system lifecycle, including how the practices identified above operate together and any cross-cutting governance, review or escalation processes. Please also indicate how often these practices are reviewed and what types of events trigger reviews and updates. Where relevant, please distinguish between different types of AI systems (e.g. generative or agentic AI).

Plan and Design

- **Practices:** Conduct Consequence Scanning workshops and Ethical Risk Assessments to define the AI's intended purpose and identify potential societal harms.
- **Tools:** Salesforce Ethical AI Maturity Model; Internal Case Review intake system.
- **Metrics:** Risk-to-Benefit Ratio; Percentage of use cases requiring human-in-the-loop oversight.

Collect and Process Data

- **Practices:** Implement strict data minimization and Zero Data Retention policies to prevent customer data from being used for unauthorized model training.
- **Tools:** Data Masking; PII (Personally Identifiable Information) Scanners; Einstein Trust Layer.
- **Metrics:** Data Lineage Completion Rate; Frequency of PII detection/redaction events.

Build, Integrate, and Interpret Model

- **Practices:** Prioritize Explainable AI (XAI) to ensure model outputs are transparent and understandable to non-technical users.
- **Tools:** Model Builder with integrated guardrails; Feature Importance Dashboards.
- **Metrics:** Model Interpretability Score; Feature Attribution weights.

Verify and Validate

- **Practices:** Execute rigorous red-teaming and adversarial testing to identify vulnerabilities in model logic and safety filters.
- **Tools:** Automated Toxicity Filters; Sandbox testing environments.
- **Metrics:** False Positive/Negative Rates in bias detection; Toxicity Thresholds (measuring the frequency of blocked harmful outputs).

Deploy

- **Practices:** Utilize staged rollouts and automated deployment pipelines that require formal governance sign-off before production.
- **Tools:** DevOps Center; Change Management Audit Logs.
- **Metrics:** Deployment Success Rate; Mean Time to Detect (MTTD) deployment-related anomalies.

Operate and Monitor

- **Practices:** Maintain continuous feedback loops and real-time observability to detect model drift or degradation over time.
- **Tools:** Einstein Trust Layer Audit Trails; Observability Dashboards.
- **Metrics:** Model Accuracy/Precision over time; Rate of Model Drift; User-reported Correctness percentages.

Please describe any measures used to ensure data quality and fairness across the data collection and curation stage (e.g., dataset audits, filtering training data to remove harmful content or personal information, fine-tuning on curated data, use of moderation tools, assessment of limitations to model generalisability).

Reviews are mandated prior to all product go-to-markets. Customer-facing policies, like the AI Acceptable Use Policy, are continuously refined as technologies and emerging risks evolve.

Salesforce ensures transparency by making every effort to include documenting training data sets within public Model Cards. To prevent harmful outputs, they utilize Model Containment Policies (system prompts) and Prompt Builder Guardrails to block sensitive topics.

8. What security measures are used in testing AI systems?

- Isolated and secure test environments
- Isolation of staging environments (from production / operational environments)
- Network isolation from external access
- Controlled environments for red-teaming and adversarial testing
- Containment measures for prompt-injection and jailbreak attempts
- Controlled security vulnerability testing

- Rate limits and resource caps
- Human-in-the-loop controls for higher-risk test cases
- Access controls, including roles-based permissions
- Audit logging and monitoring
- Use of “AI-safe” test-data, as appropriate
- Use of synthetic, anonymized, or dummy data, as appropriate

8.A. Please indicate, if relevant, the team or role responsible for overseeing secure testing and whether this responsibility is shared across multiple teams and/or functions.

Responsibility is distributed across the following key teams and roles:

- **The AI Council:** A cross-functional executive body that serves as the ultimate authority for AI governance. It brings together leaders from Legal, Security, and Product to oversee internal *Case Reviews* and approve high-risk AI agent deployments.
- **Office of Ethical and Humane Use (OEHU):** This team is responsible for the ethical dimension of secure testing. They lead Consequence Scanning workshops and define the *Ethical AI Maturity Model* used to validate that AI systems are inclusive, transparent, and safe before they move from staging to production.
- **Global Security Operations (SecOps):** The technical Security Lead role resides here, overseeing the infrastructure for Red-Teaming, adversarial testing, and vulnerability management. They ensure that test environments remain isolated and that the Einstein Trust Layer guardrails (like toxicity filtering and data masking) are functioning correctly.
- **Product and Engineering Teams:** Individual AI System Owners within specific product units (e.g., Service Cloud, Sales Cloud) are accountable for the day-to-day performance and security of their models. They are responsible for implementing the *human-in-the-loop* controls and continuous monitoring required by the AI Council.
- **Data Governance & Privacy Teams:** These teams oversee the AI-safe data lifecycle, ensuring that synthetic or anonymized datasets are used during testing and that *Zero Data Retention* policies are technically enforced.

Shared Responsibility Structure

This is not a siloed function; rather, it is an interdependent framework. While the Security team provides the *secure environment* (infrastructure), the OEHU provides the *ethical rubric* (policy), and the Product teams provide the *technical execution* (testing). All three must certify a system's readiness through the Model Registry before it is authorized for global deployment.

9. How does your organization protect intellectual property, including copyright-protected content?

9.A. Safeguards that promote lawful data use and transparency in AI training:

- Maintain a copyright policy that addresses data sourcing, content provenance, safeguards, and organizational roles and responsibilities
- Publishing “model cards” or contextual information on training data and model limitations

- Apply provenance controls to guard against training data being sourced from websites or repositories that have been determined to engage in large-scale piracy
- Conduct data cleaning, qualification, and/or tokenization activities as part of a transformative data preparation process
- Refrain from the unlawful circumvention of technological protection measures in data sourcing
- Maintain policies or practices to respect copyright holders' preferences through, where appropriate/applicable, mechanisms like Robots.txt "do not train" tags or other machine-readable protocols developed through voluntary, industry-led standards processes
- Consult and/or retain an intellectual property attorney to protect, enforce, and comply with, intellectual property rights.

9.B. Safeguards that reduce the risk of dissemination of outputs that may infringe copyright:

- Restricting outputs that raise ethical, human rights, or legal concerns—such as non-consensual intimate imagery, unauthorized deepfakes, or outputs that are substantially similar to copyrighted works.
- Filtering results using keyword or image-based restrictions to reduce risks of misuse
- Designating clear points of contact so creators or rights holders can raise concerns about infringing outputs

10. How do you protect privacy and prevent your systems from disclosing confidential or sensitive data across the AI lifecycle?

- Make information about data privacy, practices, and controls accessible to a diverse set of stakeholders
- Use security controls, including encryption at rest and in transit
- Provide organizational customers with controls to configure and implement privacy choices (e.g., data retention, logging, access, or processing settings)
- Apply Privacy-Enhancing Technologies (PETs) where applicable (e.g. differential privacy, federated learning, synthetic data generation, homomorphic encryption)

Describe key privacy measures and data-protection standards applied.

The Einstein Trust Layer guarantees Zero Data Retention, meaning it does not store any prompt or output data. Furthermore, Agentforce relies on Secure Data Retrieval, strictly enforcing permissions-based access so agents only pull data the authenticated user is authorized to see. Customers are also provided advanced security add-ons.

11. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems across the AI system lifecycle? Respondents may complete the checkbox subsections in 11.A–11.D or describe their overall approach under 11.E.

11.A. Governance, policy and assessments

- Conduct regular cybersecurity risk and vulnerability assessments tailored to AI models or systems
- Perform threat modeling that includes AI-specific risks (e.g., data poisoning, prompt injection)

- Maintain or update security policies aligned with recognized standards (e.g., NIST CSF, ISO/IEC 27001, SOC 2, CIS Critical Security Controls), and review them regularly in response to emerging risks or regulatory guidance
- Maintain vulnerability management processes (scanning, channels for external reporting, patching, remediation)
- Implement incident response, disaster recovery, or business continuity plans, including escalation procedures for high-severity incidents
- Train staff on cybersecurity best practices and AI-specific threats, including how to recognize and respond to misuse or unintended outcomes of advanced AI systems
- Integrate cybersecurity AI risks into broader cybersecurity risk management activities
- Regularly review security measures and update them in response to emerging risks, incidents, or major releases

11.B. Technical and operational controls

- Apply multi-factor authentication (MFA) and role-based access controls for advanced AI systems
- Encrypt AI assets (datasets, models, configurations) in storage and in transit as appropriate
- Restrict physical access to servers, GPUs, or devices hosting AI models, systems, and/or applications
- Monitor system activity and log access to detect anomalies or misuse
- Conduct secure AI development and/or deployment practices
- Conduct security testing focused on AI-specific risks (e.g., red-teaming, adversarial testing, stress tests)

11.C. Protection of sensitive or high value AI assets

- Limit access to proprietary AI assets based on role or need-to-know
- Use hardware security modules (HSMs) or trusted execution environments (TEEs) where feasible
- Track and log access to sensitive AI resources
- Require NDAs or confidentiality agreements for staff and contractors with access
- Store model weights, datasets, and algorithms in secure, access-controlled environments

11.D. Insider risk protection detections

- Monitor for unauthorized internal access or suspicious activity
- Conduct pre-employment screening or role-sensitive vetting
- Provide staff training on cybersecurity best practices and AI-specific threats
- Implement insider-risk escalation or response procedures

11.E. Please describe the main security measures your organization applies and indicate which are specific to particularly sensitive AI assets, where relevant.

Please see our previous responses regarding our Trust Layer.

12. What is your organization's vulnerability management process? Does your organization take actions to address identified risks and vulnerabilities, including in collaboration with other stakeholders?

- Internal vulnerability disclosure receipt and tracking processes
- Bug bounty or incentive programs for external reporters
- Investigation of reported or identified risks and vulnerabilities
- Development of patches or other mitigations for confirmed risks or vulnerabilities
- Testing of patches or other mitigations
- Scanning for and mitigation of similar risks or vulnerabilities
- Collaborative vulnerability sharing (e.g., with partners, industry groups)
- Vulnerability disclosure program that is aligned with internationally recognized standards, e.g. ISO/IEC 30111 and 29147.
- Reporting to regulators / authorities when relevant
- Recognition of external finders of confirmed risks or vulnerabilities
- Defined escalation criteria and follow-on actions for identified vulnerabilities (e.g., for external reporting or information sharing)

13. What post-deployment practices does your organization use to monitor and respond to vulnerabilities, incidents, emerging risks and misuse of advanced AI systems?

- Conduct post-deployment monitoring
- Perform root-cause analysis and apply lessons learned to improve future development and/or deployment practices
- Monitoring and emergency controls specific to agentic AI systems (e.g., ability to interrupt autonomous action sequences, override tool-use permissions, or revoke delegated authority)
- Other(s) (please specify): External Vulnerability Reporting: Salesforce operates a Responsible Disclosure Policy and a Bug Bounty Program that specifically incentivizes external researchers to report AI vulnerabilities, such as prompt injection and bias, after models have reached the public. Employee Trust Testing: Internal staff participates in an ongoing Employee Trust Tester Program to simulate real-world scenarios and identify potential vulnerabilities or jailbreak attempts that may arise post-launch.

14. Do you develop or deploy agentic AI systems (systems that can autonomously plan/act or use tools)?⁴

Yes

14.A. If yes, please indicate which categories of controls your organization applies. Please select all.

- Constraints on the agent's action space (e.g., tool-use permissions, environment sandboxing, allowlists/blocklists of resources, scope limits per task)
- Human oversight or interruptibility mechanisms (approval requirements for sensitive actions, escalation paths, interruption interfaces)
- Monitoring (anomaly detection, multi-agent interaction monitoring)
- Logging of agent actions

Please provide any further comments and implementation documentation.

Section 3 – Transparency Reporting on Advanced AI Systems

15. Does your organization publish public reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems? Are these clear and understandable by non-technical audiences?

Yes

15.A. If yes, which of the following information is included in your organization's publicly available documentation for any of your AI products?

- Risk management policies
- Summary of results from evaluations of safety, security, and societal risks including potential adverse impacts on individuals or groups (e.g. impacts on human rights, privacy or differential outcomes across user groups).
- Summary of results from red-teaming and other testing (e.g., adversarial evaluations, stress tests, and other pre-deployment safety assessments).
- Capabilities, behaviours, and limitations, including domains of appropriate and inappropriate use or indicative thresholds associated with systemic risks.
- Technical documentation and instructions for use if relevant.
- Monitoring agent to agent interactions
- Plain-language or non-technical summaries

Please provide links or references showing where this information is made public.

- **Risk management policies:** Salesforce publicly maintains its AI Acceptable Use Policy (AI AUP) and overarching Acceptable Use Policy (AUP), which establish baseline requirements for safety and ethical usage.
- **Summary of results from evaluations of safety, security, and societal risks – including risks to the enjoyment of human rights and/or risks of harmful bias, discrimination, threats to protection of privacy or personal data, or fairness:** Salesforce publishes Model Cards for its owned models, which provide structured information on ethical considerations, fairness, and safety assessments. Additionally, the organization shares high-level outcomes from its Trusted AI Review process and human rights salience assessments in public transparency reports.
- **Summary of results from red-teaming and other testing (e.g., adversarial evaluations, stress tests, and other pre-deployment safety assessments):** Model Cards include results from bias, toxicity, and other trust and safety assessments. Furthermore, Salesforce publicly details its red-teaming methodologies (such as fuzzing and manual ethical hacking) and highlights key findings through its Newsroom and annual Impact Reports.
- **Capabilities and limitations, including domains of appropriate and inappropriate use:** This information is a core component of Salesforce Model Cards, which document intended use cases and known limitations where a model may perform poorly or fail. The AI AUP also explicitly defines

prohibited high-stakes use cases, such as individualized medical or legal advice without human review.

- **Technical documentation and instructions for use if relevant:** Salesforce provides extensive documentation on its Trust website, Help & Documentation portals, and free Trailhead modules to educate users on how to deploy and interact with its AI systems responsibly.
- **Monitoring agent to agent interactions:** Salesforce has committed to testing and publishing model cards for its agentic systems and utilizes Data Cloud to provide customers with an audit trail of agent transactions for their own monitoring.

15.B. If yes, how often are such reports usually updated?

Other (e.g., triggered by material changes, incidents, or new risks)

16. How does your organization share information with a diverse set of stakeholders (other organizations, governments, civil society and academia, etc.) regarding the outcome of evaluations of risks and impacts related to an advanced AI system?

- Public documentation
- Protected information sharing
- Regulatory reporting as relevant
- Direct engagement with AI safety institutes, civil society or academic partners

Describe how evaluation results are shared and with which stakeholders (e.g. other organizations, governments, civil society and academia, etc). You may also indicate whether non-technical or plain-language summaries are provided for broader audiences.

Salesforce actively engages with a diverse set of external stakeholders to ensure transparency and accountability across the AI lifecycle:

- **Customers and Business Partners:** Evaluation results are shared via the Salesforce Compliance and Trust websites, which host Model Cards and the Agentic Benchmark for CRM to help organizations assess model reliability before deployment.
- **Governments and Intergovernmental Bodies:** Salesforce collaborates with policymakers through groups like the Hiroshima AI Process Friends Group, the NIST Center for AI Standards, and UNESCO's Global Business Council to share best practices and evaluate risk norms.
- **Civil Society and Non-Profits:** Through the Salesforce Accelerator – Agents for Impact, Salesforce provides non-profits with access to customized AI solutions and evaluation findings to drive social good.
- **Academia and Industry Research:** Salesforce participates in industry alliances like the AI Alliance, Partnership on AI, and the Data & Trust Alliance, where it shares its AI Energy Score and LLM benchmarks to advance broader academic and technical understanding of AI safety.

Plain-Language Summaries for Broader Audiences - to reach non-technical stakeholders—including the media, end-users, and the general public—Salesforce provides information in several accessible formats:

- **Nutrition-Label Inspired Model Cards:** Salesforce's Model Cards are structured similarly to nutrition labels, providing essential information about capabilities, ethical considerations, and known

limitations in a clear, standardized format.

- **Newsroom and Blogs:** Salesforce regularly publishes articles on its Newsroom that break down technical risk mitigations and agentic AI principles into understandable narratives.
- **Interactive Digital Experiences:** Resources like A Day Empowered by Agents allow broader audiences to visualize how trust guardrails and human oversight function in real-world settings without requiring technical expertise.
- **Trailhead Modules:** Free, interactive learning modules like the Trusted Agentic AI badge enable anyone to build AI literacy and understand how evaluations ensure responsible human-AI collaboration.

17. Does your organization publish public privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?

Yes

If yes, provide links to privacy policies or notices covering personal data, user inputs, and AI-generated outputs.

Salesforce Privacy Statement (General Personal Data) - This is the primary notice covering how Salesforce collects, uses, and shares personal information when you interact with its websites or services.

- **Link:** [Salesforce Privacy Information](#)
- **Coverage:** Includes data collection for service optimization, marketing, and the use of contracted service providers who may deliver AI capabilities to analyze data and create AI-generated responses.

AI Acceptable Use Policy (User Inputs & Prohibited Outputs) - This policy explicitly governs how customers and end-users may interact with Salesforce AI services, focusing on safety and ethical constraints.

- **Link:** [AI Acceptable Use Policy \(PDF\)](#)
- **User Inputs:** Mandates that customers disclose to end-users when they are interacting with automated systems like Einstein bots or Agentforce agents.
- **Generated Outputs:** Prohibits the use of AI to generate individualized advice in high-stakes areas (e.g., medical, legal, or financial) without a human in the loop, and bans the creation of deceptive or harmful content.

Salesforce Compliance Website (Model Cards) - For specific details on what data was used to train individual models and known limitations, Salesforce publishes standardized documentation.

- **Link:** [Salesforce Compliance - AI Documents](#)
- **Coverage:** Hosts Model Cards that detail training data sources, evaluation metrics, and intended vs. inappropriate use cases for Salesforce-owned models.

18. How does your organization provide information about the sources of data used for the training of advanced AI systems, as appropriate, including information related to the sourcing of data annotation and enrichment?

- Public documentation
- Provide information or tools to downstream actors on data quality
- Regulatory reporting as relevant
- Direct engagement with civil society or academic partners

Please provide links to any publicly available information.

We have previously provided this information above.

19. Does your organization demonstrate transparency related to advanced AI systems through any other methods?

Yes

If yes, describe additional method(s), including where relevant transparency measures related to AI agents

Embedded Technical Transparency: The Einstein Trust Layer - The Einstein Trust Layer acts as a security blanket that provides real-time transparency into how data is handled during AI processing.

- **Comprehensive Audit Trails:** It maintains detailed logs that record how data is used and processed by AI models, allowing organizations to audit agent decisions.
- **Dynamic Data Masking:** The system visually masks sensitive information (like SSNs) during the processing phase and unmask it only when returned to the user, ensuring the how of data protection is visible to administrators.
- **Toxicity Scoring:** The Trust Layer scans prompts and responses for harmful content, providing confidence scores that allow users to see why certain content was flagged or mitigated.

Explainable AI (XAI) Tools - Salesforce provides a suite of tools designed to peek into the black box of AI decision-making:

- **Feature Importance:** Highlights exactly which influential factors led a model to a specific prediction or decision.
- **Visual Decision Trees:** Maps out the logical steps an AI model took to arrive at a particular outcome, making the reasoning understandable for non-technical users.

Interactive Transparency Experiences - To help stakeholders understand complex agentic behaviors, Salesforce uses interactive and multimedia formats:

- **A Day Empowered by Agents:** An interactive digital experience that allows users to visualize how trust guardrails and human oversight function within real-world real-time workflows.
- **Salesforce+ Content:** Dedicated video series and keynotes, such as the Security Keynote at Dreamforce, provide in-depth walkthroughs of the company's AI safety and transparency work for the public.

Educational & Literacy Initiatives - Transparency is fostered through AI literacy, empowering users to identify and mitigate risks themselves:

- **Trailhead Learning Modules:** Interactive badges like the Trusted Agentic AI module educate over 111,000 users on the specific trust patterns (such as intentional orchestration) used in Salesforce

products.

- **Public Best Practices:** Salesforce regularly publishes Trust Patterns (e.g., standardized disclaimers and human-AI handoff protocols) to guide customers on making their own AI implementations transparent to their end-users.

Social & Ethical Reporting

- **Human Rights Salience Assessments:** Salesforce conducts and shares findings from company-wide assessments to show how its AI systems align with international human rights standards.
- **AI Sustainability Analytics:** Through tools like the AI Energy Score, Salesforce provides transparency into the environmental impact of AI models, detailing energy consumption and carbon footprints.

20. Does your organization publish statistical insights derived from anonymised, aggregated usage data to advance understanding of AI adoption and its economic impacts in areas including occupations and tasks by region, sector and country? To this end, do you contribute to international cooperation efforts?

Yes

If yes, please provide links or describe the scope of insights published, including where relevant contributions to understanding AI adoption in specific sectors (e.g. public administration, education systems).

Salesforce publishes statistical insights that track the evolution of the AI economy and workforce adoption, derived from aggregated data and global surveys of employees, customers, and partners:

- **Workforce & Industry Adoption Trends:**
 - **Slack Workforce Index:** Salesforce publishes data on AI adoption among global desk workers, highlighting usage growth (e.g., a 233% increase in usage over six months) and the impact of AI on task efficiency and job satisfaction.
 - **State of the AI Connected Customer:** This annual report, informed by over 16,500 voices, provides cross-country and cross-sector insights into customer trust, expectations for automated agents, and the shift from predictive to agentic AI.
- **Economic Impact Assessments:**
 - **The Salesforce Economy Study:** In partnership with IDC, Salesforce publishes aggregated data on how AI will create millions of net-new jobs (estimated 11.6 million by 2028) and trillions in business revenue across specific regions such as the U.S., Japan, Germany, and France.
 - **AI Energy Score:** Salesforce provides statistical benchmarks for the environmental impact of AI models, enabling developers to compare energy consumption and carbon footprints globally.
- **International Cooperation Efforts:**
 - **G7 Hiroshima AI Process (HAIP):** Salesforce was one of the first companies to submit a transparency report under this framework to coordinate international norms on AI safety and disclosure.

- **NIST & OECD:** Salesforce actively contributes to the NIST Center for AI Standards and the OECD AI Principles to share best practices for identifying and mitigating socio-economic risks.

Public Documentation Links:

- [Salesforce Newsroom \(Research & Stories\)](#)
- [Salesforce Compliance & AI Documentation](#)
- [State of the Connected Customer Report](#)

Please estimate the total staff time and annual expenses associated with conducting compliance activities and include any further comments and implementation documentation (optional).

No answer provided

Section 4 – Organizational Governance, Incident Management, and Transparency

21. How has AI risk management been embedded in your organization's governance framework?

- Board of Directors oversight
- Executive oversight
- Senior management oversight
- Governance council or committee
- Independent advisory body
- Other (e.g., distributed / multi-institutional governance arrangement) (please specify): Human Rights Steering Committee, which includes executives across legal, privacy, and procurement to oversee human rights risks.

Please describe how AI risk management is integrated into governance structures, including: which roles or bodies have advisory versus decision-making authority; which roles or bodies have final accountability for AI risk-related decisions; and how often policies are reviewed or updated (e.g., following major incidents, regulatory changes, or material updates to systems).

Salesforce integrates AI risk management using a top-down, coordinated approach that ensures ethical considerations are embedded into both strategy and operational workflows:

- **Decision-Making Authority:** Decision-making is shared across specialized bodies to balance speed with safety. The AI Trust Council (senior executives across Security, Product, Legal, etc.) aligns and accelerates product decisions. While these councils have advisory and decision-making roles, the executive leadership team and the Board's Cybersecurity and Privacy Committee provide final review and approval for overarching strategies.
- **Final Accountability:** Final accountability for AI risk-related decisions rests with the Cybersecurity and Privacy Committee of the Board of Directors. This committee meets with the Chief Ethical and Humane Use Officer and other management to steer trusted AI priorities.
- **Review and Update Cadence:** Policies are reviewed following major incidents, significant regulatory changes, or material updates to AI systems. For example, the AI Acceptable Use Policy

(AI AUP) was updated in 2024 to specifically address risks associated with agentic AI. Internal reviews for AI use cases are mandatory prior to any go-to-market release.

22. Are relevant staff trained on your organization's governance policies and risk management practices?

Yes

If yes, describe training methods (e.g., onboarding modules, regular sessions, and confidential channels to report concerns).

Salesforce utilizes a variety of internal programs to embed its Trusted AI Principles into the workforce culture:

- **Continuous Learning & Reskilling:**
 - **Trailhead for Employees:** Salesforce leverages its free online learning platform, Trailhead, to offer personalized, role-specific modules like the Trusted Agentic AI badge. Over 111,000 users, including internal staff, have earned this credential to build AI literacy.
 - **Onboarding and Professional Development:** New hires receive comprehensive training on ethical use policies, while existing staff have access to tuition reimbursement and leadership development initiatives focused on the agentic era.
- **Specialized Hands-on Training:**
 - **Employee Trust Tester Program:** This initiative creates specific opportunities for employees to inform product development by identifying gaps and vulnerabilities in AI technology. It upskills employees on Agentforce while providing data to inform business decisions.
 - **Technical Upskilling:** Underrepresented students and staff in computer science gain access to hands-on training for building and using AI agents through the Futureforce Tech Launchpad.
- **Employee Feedback and Governance Engagement:**
 - **Reverse Mentorship Program:** Senior leaders are paired with entry-level employees to share insights on emerging tech and workplace culture, fostering a top-down and bottom-up flow of governance knowledge.
 - **Internal AI Governance Review:** Multidisciplinary teams from Business Technology, Security, Legal, and Ethics are trained to facilitate the swift but responsible deployment of internal AI tools.
- **Confidential Reporting Channels:**
 - **Policy Operations:** Salesforce maintains a Policy Operations function that manages case management and content safety operations. This team is responsible for responding to internal and external user reports and escalations related to ethics or policy concerns.
 - **Compliance and Safety Reporting:** Staff are provided with confidential channels to report potential safety, ethical, or policy violations to the Office of Ethical and Humane Use and the Legal team.

23. Are steps taken to address reported incidents documented and maintained internally?

Yes

If yes, describe how incident records are kept and used for follow-up (e.g., incident logs, root-cause analysis reports, internal tracking systems). You may indicate whether and how you define escalation thresholds for when incidents must be surfaced to senior leadership or external parties.

Salesforce utilizes a coordinated system to track, manage, and learn from incidents across its AI lifecycle:

- **Internal Tracking Systems:** The Policy Operations function within the Office of Ethical and Humane Use manages specialized case management and content safety operations. These operations include responding to user reports and maintaining detailed documentation of incident timelines and outcomes.
- **Incident Logs and Audit Trails:** Technical safeguards, such as the Einstein Trust Layer, automatically log toxicity scores and interaction data into an audit trail stored in Data 360. These logs visualize toxicity trends over time, providing a verifiable record of system behavior for internal review.
- **Root-Cause Analysis:** Salesforce's incident response team uses Agentforce to automatically triage next steps for security incidents and evaluate the quality of responses. This process includes conducting post-incident analysis to identify the root cause of an issue and areas where technical or non-technical mitigations (such as updated Trust Patterns) are required.
- **Escalation Thresholds:**
 - **Strategic Escalation:** Salesforce defines clear paths for surfacing high-impact incidents to senior leadership, including the Cybersecurity and Privacy Committee of the Board of Directors and the AI Trust Council.
 - **Automated Thresholds:** Escalation is often encoded directly into the system; for example, high-stakes situations involving regulatory compliance, significant financial exposure, or agent uncertainty automatically trigger a transfer to human review.
 - **Policy Escalation:** Concerns related to high-priority ethical use cases follow a structured process where recommendations are presented to executive leadership for final review and approval.

24. Does your organization share research and best practices on addressing or managing risk?

Yes

If yes, describe how research or good practices are shared (e.g., publications, conferences, collaborations with other sectors).

- **Academic and Peer-Reviewed Publications:** The Salesforce AI Research team regularly publishes state-of-the-art research on trust and safety metrics, including emergent failures like echoing in agent-agent interactions and robust reasoning frameworks. The company has also open-sourced safety libraries, such as AuditNLG, to assist the broader community in identifying harmful model outputs.
- **Global Conferences and Summits:** Salesforce is a consistent participant in major international convenings, including the AI Action Summit and Bletchley Park AI Safety Summits. At Dreamforce, the company hosts dozens of sessions specifically dedicated to responsible AI, agent safety, and human-AI collaboration.

- **Intergovernmental and Industry Partnerships:** The organization collaborates with global policymakers through groups such as the Hiroshima AI Process Friends Group, the NIST Center for AI Standards, and UNESCO's Global Business Council. Salesforce is also an active member of the AI Alliance, Partnership on AI, and the Data & Trust Alliance, working to develop sector-wide standards for data integrity and bias mitigation. We also convene our own workshops across industry, academia, NGO, civil society, etc. for knowledge sharing
- **Public Documentation and Interactive Digital Resources:** Best practices are shared via the Salesforce Newsroom, which published over 10 blog posts this year on ethical AI use and agentic AI risks. Salesforce also launched A Day Empowered by Agents, an interactive experience designed to guide leaders in navigating responsible agent deployment.
- **Collaborative Safety Tools:** In partnership with Hugging Face, Cohere, and Carnegie Mellon University, Salesforce released the AI Energy Score to standardize reporting on the environmental impact of AI models.
- **Free Public Learning (Trailhead):** Over 111,000 users have completed Salesforce's Trusted Agentic AI module, which teaches the core trust patterns and best practices for managing AI autonomy.

25. Does your organization use international technical standards or best practices for AI risk management and governance policies?

Yes

If yes, please indicate which standards or frameworks apply:

- ISO/IEC standards (e.g., ISO/IEC 27001, ISO/IEC 42001)
- NIST AI Risk Management Framework (AI RMF)
- OECD AI Principles
- Industry-led best practices or codes of conduct
- Other international or regional standards (please specify): OWASP for LLM

If yes, describe which international technical standards or best practices your organization uses.

NIST AI Risk Management Framework (AI RMF)

Salesforce aligns its AI lifecycle with the NIST AI RMF 1.0. We specifically map our internal processes to the framework's four core functions:

- **Govern:** Establishing a risk-aware culture through our cross-functional AI Council.
- **Map:** Contextualizing AI systems within their specific business and ethical environments.
- **Measure:** Using quantitative metrics to assess bias, toxicity, and model drift.
- **Manage:** Implementing technical guardrails, such as the Einstein Trust Layer, to mitigate identified risks.

ISO/IEC Standards

To maintain a certifiable and auditable governance structure, we reference:

- **ISO/IEC 42001 (AI Management System):** We use this as the foundational management-system backbone to demonstrate governance maturity to auditors and enterprise partners.

- **ISO/IEC 23894:** We utilize this standard's guidance on AI-related risk management to refine our internal Ethical AI Maturity Model.

OECD AI Principles

Salesforce's Five Trusted AI Principles (Responsible, Accountable, Transparent, Empowering, and Inclusive) are explicitly mapped to the OECD Principles on Artificial Intelligence. This ensures our corporate values support international priorities regarding human rights, diversity, and fair distribution of AI benefits.

Regulatory Frameworks & Regional Best Practices

We actively adapt our technical controls to meet the requirements of emerging sovereign laws:

- **EU AI Act:** We have aligned our *Model Registry* and documentation standards with the EU's risk-based classification system, ensuring high-risk systems meet Article 13 requirements for transparency and logging.
- **G7 Hiroshima AI Process:** Salesforce follows the International Code of Conduct for Organizations Developing Advanced AI Systems, focusing on post-market monitoring and vulnerability disclosure.
- **techUK and Regional Roadmaps:** We participate in national-level best practice development, such as the UK's AI Safety Summit initiatives and Southeast Asia's *Penta-helix* collaborative models.

Industry Standards

- **OWASP for LLMs:** We integrate the OWASP Top 10 for Large Language Models into our security testing to defend against specific vulnerabilities like prompt injection and insecure output handling.
- **Model Cards:** We utilize the industry-standard format to provide transparent disclosures on training data, intended use cases, and model limitations.

Any further comments and implementation documentation.

No answer provided

Section 5 – Content Authentication & Provenance Mechanisms

26. What mechanisms, if any, does your organization put in place to allow end-users, where possible and appropriate, to know when they are interacting with an advanced AI system developed by your organization?

- User disclosure (e.g., text notice, verbal statement, or system label)
- Persistent visual or audio indicators (e.g., watermark, icon, chime, banner) during AI interaction
- Distinct branding or labeling for AI-generated content (e.g., "AI-generated" tag, metadata label)
- Documentation or FAQs explaining when and how AI is used in the system
- Access to provenance information (e.g., model version, content authenticity signals)
- Disclosure to end-users when an AI system is acting autonomously or executing actions on their behalf (e.g., making purchases, sending communications, modifying files)

Provide any additional details or context on how or when users are informed they are interacting with advanced AI systems (e.g. if such mechanisms are planned, in testing, or applied only to certain

products).

- **User disclosure (e.g., text notice, verbal statement, or system label):** Salesforce's AI Acceptable Use Policy (AI AUP) mandates that customers disclose to their users when they are interacting with an AI feature, such as Agentforce agents. Standardized Trust Patterns also require disclaimers to disclose AI-generated content to both internal and external recipients.
- **Persistent visual or audio indicators (e.g., watermark, icon, chime, banner) during AI interaction:** UI indicators like different profile pictures for agents are recommended to distinguish them from human users.
- **Distinct branding or labeling for AI-generated content (e.g., AI-generated tag, metadata label):** Salesforce implements standardized labeling for autonomously delivered AI content to ensure its origin is clear.
- **Documentation or FAQs explaining when and how AI is used in the system:** Salesforce provides extensive documentation, including Model Cards that explain model capabilities and intended use cases, and public FAQ resources.
- **Access to provenance information (e.g., model version, content authenticity signals):** Through Citations and Audit Trails, Salesforce provides end-users with traceability and explainability, allowing them to see exactly what data was used in a generation.
- **Disclosure to end-users when an AI system is acting autonomously or executing actions on their behalf:** Policy requirements ensure users are informed at the start and at regular intervals when an AI system is acting autonomously.

Salesforce focuses on Transparency by Design to ensure that disclosure is integrated into the user experience rather than being an afterthought.

- **Mindful Friction:** Salesforce designs user interfaces with equal visual weight for actions like Send and Edit to encourage users to review AI-generated content before execution.
- **Human-AI Handoffs:** Standardized patterns ensure that high-impact tasks include human checkpoints and clear escalation paths to a human professional if the AI encounters uncertainty.
- **Accessibility Integration:** Salesforce utilizes Agentforce usability studies with users with disabilities to refine how AI structure (such as message alignment and labels) is communicated via assistive technology like screen readers.

27. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems?

Yes

27.A. If yes, how?

- Latent disclosure (e.g., hidden provenance information)
- Manifest disclosure (e.g., visible label)
- Detection tool(s) for hidden provenance signals (e.g., watermark or metadata checks)
- Use of international or industry standards (e.g., C2PA)

27.B. If yes, does your organization use international technical standards or best practices when developing or implementing content provenance?

Yes

27.C. Does your organization collaborate on or invest in research to advance content authentication and provenance?

Yes

Describe how provenance mechanisms are implemented and which standards or tools are used, and any collaborations or investments to improve content authentication (e.g., C2PA credentials, joint R&D projects, open-source initiatives, standard-setting groups). You may also indicate whether these mechanisms are tested for robustness against removal, tampering, or circumvention.

Salesforce operationalizes its provenance strategy through Transparency by Design, ensuring that authenticity mechanisms are embedded throughout the AI lifecycle:

- **Implementation & Tools:** The primary mechanism is the Einstein Trust Layer, which acts as a gateway for all AI calls. It creates a cryptographically secure Audit Trail and provides Citations that link AI outputs back to original data sources in Data Cloud. This ensures that users can trace exactly what data influenced a specific generation.
- **Collaborations:** Salesforce is an active member of the AI Alliance and the Data & Trust Alliance, where it works with peers to develop Content Credentials—tamper-evident metadata that functions like a nutrition label for digital content.
- **Testing for Robustness:** These mechanisms are stress-tested through Ethical Red Teaming and automated fuzzing to identify vulnerabilities where provenance metadata might be removed or circumvented. For example, Salesforce tests its system prompts and Model Containment Policies to ensure that safety instructions (which include disclosure requirements) cannot be bypassed by malicious actors.

Any further comments and implementation documentation.

No answer provided

Section 6 – Research & Investment to Advance AI Safety & Mitigate Societal Risks

28. Does your organization advance research and investment in the following areas?

- Security
- Safety
- Data quality and representativeness
- Information integrity
- System performance and outcomes (e.g. accuracy, consistency)
- Explainability or interpretability
- Transparency
- Robustness or reliability

If yes, please describe key research or investment initiatives in the areas selected above (e.g., dedicated safety teams, funding of research programs, contributions to open-source safety tools, publishing research, partnerships and collaborative research initiatives). You may also indicate whether related outputs are publicly shared, shared with partners or within consortia or not publicly disclosed.

Salesforce operationalizes these priorities through several specialized initiatives and collaborative efforts:

- **Dedicated Ethical Teams:** The Office of Ethical and Humane Use guides responsible design, while the Testing, Evaluation, and Alignment team focuses on rigorous manual and automated red teaming to uphold safety standards.
- **Academic and Peer-Reviewed Research:** Salesforce regularly publishes work on trust and safety metrics, including the creation of the first-ever Agentic Benchmark for CRM to evaluate privacy, truthfulness, and robustness in enterprise contexts.
- **Open-Source and Public Tools:** Salesforce collaborates with partners like Hugging Face and Carnegie Mellon University to develop and release tools such as the AI Energy Score, which standardizes environmental impact reporting for AI models.
- **Publishing and Transparency:** The company has committed to testing and publishing Model Cards for all Salesforce-owned models. These nutrition-label-inspired documents publicly disclose training data provenance, limitations, and results from bias and toxicity assessments.
- **Global Collaborative Initiatives:** Salesforce actively participates in intergovernmental and industry alliances, including the G7 Hiroshima AI Process, the AI Alliance, and the Data & Trust Alliance, to develop globally applicable risk-based norms and standards.
- **Internal Literacy and Testing:** The Employee Trust Tester Program and Trailhead modules (e.g., the Trusted Agentic AI badge) serve as internal investments to upskill staff and identify potential system vulnerabilities before they reach customers.

Sharing Status: Related outputs such as Model Cards, Agentic Benchmarks, and specialized research blogs are publicly shared via Salesforce's Compliance and Trust websites to drive industry-wide accountability.

29. Does your organization pursue research or investment related to socio-economic or environmental considerations in the development and deployment of AI?

- Socio-economic considerations (e.g., AI adoption and productivity, workforce transitions, skills development, access to AI tools)
- Environmental considerations (e.g., energy use, computational efficiency, resource use)

If yes, please provide examples of relevant research or investment initiatives for the areas selected above.

Socio-Economic Risk Mitigation — Salesforce invests in global programs to address the AI skills gap and ensure AI benefits reach underserved communities:

- **AI for All & Trailhead:** In September 2024, Salesforce committed \$50+ million, waiving fees on premium AI courses, certifications, and instructor-led credential programs on Trailhead through the

end of 2025, and opened a dedicated AI training hub at its San Francisco HQ. Since June 2023, Trailhead learners have earned more than 2.6 million AI and data badges, and the platform serves over 11 million learners worldwide building AI, generative AI, responsible AI, and data-literacy skills.

- **Global AI Skilling Commitments:** In November 2025, Salesforce partnered with SmartBridge, IndiaAI, and India's Ministry of Electronics and IT on *Yuva AI Bharat: GenAI Skill Catalyst*, committing to train 100,000 learners (emphasizing Tier 2/Tier 3 cities) by end of 2026 via Trailhead.
- **Philanthropic Giving:** Salesforce's education program has delivered more than \$50 million to local schools to date, part of its broader philanthropic model that gives 1% of product, equity, and employee time.
- **Futureforce Tech Launchpad:** A 10-week program giving underrepresented computer-science students technical training on building and using AI agents.
- **Salesforce Ventures Impact Fund:** Invests in enterprise software companies driving measurable social impact, focused on Education & Workforce Development and Financial Inclusion.

Environmental Risk Mitigation — Salesforce treats sustainable AI as a business imperative:

- **AI Energy Score:** Developed with Hugging Face and Carnegie Mellon and unveiled at the AI Action Summit, this benchmarking tool lets developers evaluate and compare AI model energy consumption to drive market preference toward efficient models. A related Salesforce survey found 65% of sustainability professionals say their companies must balance AI's benefits against its environmental impact, amid projections that AI-driven data-center power use will double by 2026.
- **Right-sized AI design:** Agentforce is built to be *optimised out-of-the-box*, using Data Cloud/Platform metadata for accuracy while minimizing wasted computation, avoiding the need for energy-intensive custom model training per customer.
- **Agentforce for Net Zero:** Combines emissions management, reporting, and forecasting with an agentic AI layer that lets sustainability teams query emissions/water/waste data in natural language, surface efficiency opportunities, and auto-generate framework-aligned reports

Any further comments and implementation documentation.

No answer provided

Section 7 – Advancing Human and Global Interests

30. Does your organization support any digital literacy, education or training initiatives to improve user awareness and/or help people understand the nature, capabilities and limitations of advanced AI systems?

Yes

If yes, describe any digital literacy or training initiatives your organization supports (e.g., user education, public AI-education programs, school or university partnerships, community workshops). You may also indicate whether these efforts include or prioritise underserved or underrepresented communities.

Salesforce supports a range of digital literacy and AI-education programs spanning user training, public education, and community/academic partnerships — several with an explicit focus on underserved and underrepresented groups.

- **Trailhead (core learning platform):** Salesforce's free online learning platform now serves over 11 million learners worldwide, who have earned more than 2.6 million AI and data-literacy badges since June 2023, covering AI, generative AI, responsible AI, and data skills.
- **AI for All initiative:** Announced September 2024, Salesforce committed \$50+ million to waive fees on premium AI courses, certifications, and instructor-led credential programs on Trailhead (through end of 2025), and opened a dedicated AI training hub at its San Francisco headquarters to expand public access to AI skilling.
- **Global skilling partnerships (underserved-community focus):** In November 2025, Salesforce partnered with SmartBridge, IndiaAI, and India's Ministry of Electronics and IT on *Yuva AI Bharat: GenAI Skill Catalyst*, committing to train 100,000 learners via Trailhead by end of 2026, with explicit emphasis on reaching Tier 2 and Tier 3 cities — i.e., prioritizing communities with less access to tech-sector opportunity.
- **School partnerships and giving:** Salesforce's education program has delivered more than \$50 million to local schools to date, part of the company's broader 1:1:1 philanthropic model (1% of product, equity, and employee time), which underpins its education and workforce-development giving.
- **Futureforce Tech Launchpad:** A 10-week program providing underrepresented computer-science students with hands-on technical training in building and using AI agents — a direct pipeline effort aimed at diversifying the AI/tech talent pool.

31. Does your organization pursue AI projects that aim to deliver societal or public benefits, including through partnerships with civil society and community groups?

Yes

If yes, please provide examples, including any collaborations with civil society, community groups, or efforts to address broader societal challenges.

Salesforce aligns its technology, philanthropy, and community partnerships with the UN's 17 SDGs, and is increasingly channeling its AI capabilities specifically toward this work — including through direct partnerships with civil society and nonprofit organizations.

- **Salesforce Accelerator — Agents for Impact:** Provides technology, funding, and pro bono expertise to help nonprofits build and customize AI agents, including a cohort focused exclusively on nature-related work — directly applying AI to civil-society-led SDG projects (e.g., climate action, SDG 13/15).
- **Salesforce.org — Technology for Social Change:** Equips nonprofit, educational, and philanthropic organizations with CRM and AI-enabled technology to operate more efficiently, raise funds, and connect with communities, alongside training and partnership-building support.

- **10 Priority SDGs framework:** Salesforce has identified 10 priority SDGs it works to advance through its technology, philanthropy, equality/sustainability initiatives, and partnerships with community organizations — explicitly committing to listening and learning from its nonprofit partners in shaping that work.
- **1-1-1 philanthropic model / Pledge 1%:** As a founding partner of Pledge 1%, Salesforce's 1-1-1 model (1% of product, equity, and employee time) has driven \$676M in all-time giving and 8.5 million employee service hours, supporting 56,000 nonprofits and schools that use Salesforce technology — directly tied to SDG 17 (Partnerships for the Goals).
- **Education and workforce development (SDG 4 & 8):** Salesforce has given over \$90 million to schools worldwide (SDG 4: Quality Education) and supports workforce-development grants and partnerships (SDG 8: Decent Work and Economic Growth).
- **#TeamEarth:** A global campaign aligning Salesforce's sustainability and accessibility goals with the SDGs, aiming to mobilize businesses and individuals toward a fairer, more equal, and sustainable future.

Any further comments and implementation documentation.

No answer provided