Salesforce
Trusted AI and Agents Impact Report

February 2025

Introduction

Six years ago, Salesforce established the **Office of Ethical and Humane Use**, emphasizing the strategic business imperative of ethics and human rights in business. The Office was founded with the charter to act as a central body that guides the design, development, and use of Salesforce technologies to ensure positive impacts and mitigate harms.

Since then, our world and our technologies have only gotten more advanced and complex. As artificial intelligence continues to rapidly evolve and integrate into our work lives, ensuring its trustworthiness is a business imperative – especially in the era of AI agents, which Salesforce is defining and leading through Agentforce.

At Salesforce, Trust is our #1 core value and we have long been committed to developing and deploying AI solutions that are safe, secure, and trustworthy. While trust is crucial for both consumer and enterprise AI, the stakes are significantly higher in the enterprise realm. Enterprise AI has the potential to impact entire organizations, including their employees, customers, and even entire industries, whereas consumer AI typically affects individuals. Despite sometimes being overlooked in AI discussions, enterprise AI is poised to be the driving force in transforming work. Our approach at Salesforce is centered on empowering our customers, enabling organizations of any size to reimagine their business with AI, including by bringing humans and agents / AI together to deliver customer and stakeholder success.

In this first-ever Trusted AI Impact Report, we aim to reinforce that commitment and provide a comprehensive overview of our efforts and learnings on this topic.

The report will cover the foundational principles, policies, and decision-making structures that guide our AI initiatives across the company and for our customers in a trusted manner. This includes deep dives on our responsible agentic AI principles, AI Acceptable Use Policy, and Ethical Use Advisory Council, among other structures we have in place to ensure the responsible and trusted creation and use of AI. We will share our responsible AI review, testing, and mitigation process, presenting case studies to illustrate the practical application of our principles in product design and development in partnership with teams across the company. And, we'll focus on the broader impact of our AI initiatives and our role within the AI ecosystem from our employees, to global government stakeholders, to the health of our planet, and more.

By sharing our learnings and showing our work, we aim to empower and enable others to join us in building AI solutions that are not only innovative but also ethical and trustworthy.

Sabastian Niles
President and Chief Legal Officer, Salesforce

Paula Goldman
EVP, Chief Ethical and Humane Use Officer, Salesforce

This report includes deep dives on our responsible agentic AI principles, AI Acceptable Use Policy, and Ethical Use Advisory Council, among other structures we have in place to ensure the responsible and trusted creation and use of AI.

Contents

Trusted Agents Principles and Decision-Making Structures

Trusted Agentic AI Guiding Principles

Our **Office of Ethical and Humane Use** has been thinking about the risks and opportunities for AI for more than five years now, through the eras of predictive and generative AI, and now for agentic AI. We developed our first set of **trusted AI principles** in 2018 before the widespread adoption of generative or agentic AI and we continue to guide the responsible development and deployment of AI here at Salesforce through principles, policies, and products.

As we entered into the era of generative AI in early 2023, we augmented our trusted AI principles with a set of **5 guiding principles for developing responsible generative AI** as the first enterprise company to put out guidelines in this emerging space. **These principles** still hold true for the third era of AI we find ourselves in now – the era of AI agents.

Accuracy

Safety

Honesty

Empowerment

Sustainability

See full explanation on following page.

As Agentforce continues to evolve, we're focused on **intentional design and system-level controls** that enable humans and AI to work together successfully – and responsibly. By adhering to these principles and guidelines, Salesforce is committed to developing AI that is not only powerful and efficient but also ethical and trustworthy. We believe that by focusing on these core principles, we can build AI solutions that our customers can trust and rely on, paving the way for a future where humans and AI work together seamlessly and responsibly.

Accuracy

Prioritize accuracy in Agent outputs and results. Thoughtful topic and constraints in setup advance accuracy with clear instructions on what actions the agent can and can't take on behalf of a human. And, if there is uncertainty about the accuracy of its response, citations, explainability, or other means can be leveraged by the agent to enable users to validate these responses.

Safety

Mitigate bias, toxicity, and harmful outputs by conducting bias, explainability, and robustness assessments, and ethical red teaming. Prioritizing privacy protection in agent responses and actions for any personally identifying information (PII) present in the data used for training and creating guardrails can prevent additional harm.

Honesty

Respect data provenance and ensure consent to use data (e.g., open-source, user-provided) when collecting data to train and evaluate models. Additionally, transparency by design is key – when content is autonomously delivered (e.g., chatbot response to a consumer, use of watermarks), it is important to be transparent that an AI has created content.

Empowerment

Prioritize the human-AI partnership and design meaningful and effective hand-offs. There are some cases where it is best to fully automate processes but there are other cases where AI should play a supporting role to the human – or where human judgment is required. Identify the appropriate balance to "supercharge" human capabilities and make these solutions accessible to all.

Sustainability

Create right-sized models where possible to reduce carbon footprint. When it comes to AI models, larger doesn't always mean better: In some instances, smaller, better-trained models outperform larger, general-purpose models. Additionally, efficient hardware and low-carbon data centers can further reduce environmental impact.

AI Governance

In addition to principles and policies, having the right decision-making structures in place is critical to ensuring the responsible development and deployment of AI. At Salesforce, there are many governance and accountability structures for trusted AI. To note a few:

- **Salesforce's Cybersecurity and Privacy Committee of the Board of Directors (quarterly)** meets with our Chief Ethical and Humane Use Officer quarterly to receive updates and provide feedback on key trusted AI priorities.
- **Senior leadership engagement (as needed):** The Office of Ethical and Humane Use also has regular interactions with the executive leadership team of the company to discuss policy and product topics for review and approval.
- **Human Rights Steering Committee (quarterly):** The Committee includes executives from the Office of Ethical and Humane Use, Legal, Privacy, Employee Success, Procurement, Government Affairs, Equality and Sustainability who jointly oversee our human rights program, including efforts to monitor, identify and mitigate salient human rights risks.
- **The AI Trust Council (bi-weekly):** Deeper dive: Comprised of executives across Security, Product, Engineering, AI Research, Product Marketing, Legal, UX, and Ethical and Humane Use, the AI Trust Council was formed to align and speed up decision-making for AI products
- **Ethical Use Advisory Council (quarterly):** Deeper dive: The Ethical Use Advisory Council is our overarching body that guides the Office of Ethical and Humane Use in its product and policy recommendations to leadership. This Advisory Council was established in 2018 and is composed of external experts from academia and civil society along with internal VP+ level executives and frontline employees (below VP level). The Advisory Council provides strategic guidance, feedback, and counsel on the top priorities of the Office of Ethical Use.

Ethical Use Policy

Ethical Product Use Frameworks

At Salesforce, we recognize that new technologies like **generative AI** and agentic AI require thoughtful implementation and guidance on how our customers can and can't use our products.

To this end, we stood up the **Ethical Use Policy team** within our **Office of Ethical and Humane Use** in 2019, which conducts rigorous research and engages with various internal and external stakeholders to carefully consider the impact of AI product use cases, their potential connection to harm, effectiveness of proposed actions, and to alignment with our guiding principles:

Human Rights. We work to ensure the direct use of our technologies upholds equal and inalienable protections for all individuals.

Privacy. We advance privacy best practices in our product design, empowering customers to protect individuals' data.

Safety. We aim to protect people from direct harm from the use of our technology.

Transparency. We work to ensure that our models and features respect data provenance and are grounded in your data whenever possible. We strive for transparency in our autonomously delivered AI content.

Inclusion. We promote equal access to technology, creating opportunities for all.

The Ethical Use Policy team is comprised of three functions working together to develop and operationalize policies and guidelines that promote the responsible use of our services:

Policy development: Developing and implementing customer-facing policies,

Strategic response: Risk management and crisis response, and

Policy operations: Tooling, case management, content safety

Ethical Use Policy Development

Guided by our principles, our Policy team leverages our internal case process to listen, identify, analyze, recommend and implement policy decisions. The team has established a rigorous, structured, and repeatable process for ethical use case concerns of high priority and urgency:

1 Identification

Identify the area where a policy is needed or an existing policy needs refinement.

2 Analysis

Research and analyze third party policies, industry standards, regulations related to the policy area

3 Stakeholder Engagement

Informed by multi-stakeholder input throughout the company and within the industry at large

4 Advisory Council Engagement

Engage with our ethical use advisory council to ensure our policies are on the right path.

5 Leadership Recommendation

Make a recommendation on what the policy should be to leadership and make changes based on their input.

6 Customer Conversations

Work with our customers to prioritize their needs.

7 Implementation

Enforcement of the policy to ensure we can take necessary action.

8 Ongoing Learning

Policies are living documents. It's never one and done, ensure we can make changes as needed.

Ethical Use Policy Development

We recognize the importance of collaboration in this crucial work. That's why we actively listen to our employees, partners, customers, impacted communities, and experts. For each issue, we carefully consider the direct product use case, its potential connection to harm, alignment with our guiding principles, effectiveness of proposed actions, and the current geopolitical landscape. This ensures our policies are informed by multiple perspectives, our guiding principles, and Salesforce's core values.

These policies take shape in the form of our customer-facing **Acceptable Use Policy (AUP)** and **AI Acceptable Use Policy (AI AUP)**, which are updated in partnership with our Legal team.

The AI AUP was launched in 2023 as its own policy specifically to cover the use of generative AI – to align with industry standards, applicable regulations, our third party LLM partners' policies, and to protect our customers and communities. It has been updated and clarified for agentic AI in 2024. These policies allow our customers to use Salesforce products with confidence, knowing they and their end users are receiving a truly ethical AI experience from product development to deployment. Let's take a look at what that looks like in practice.

Policy Case Studies

CASE STUDY 1

Facial Recognition

The Decision: In 2017, Salesforce evaluated the risks associated with facial recognition, including concerns about accuracy, bias, and privacy.

Informed by ongoing consultations with customers, privacy advocates, and policymakers, Salesforce made the deliberate choice **not to develop facial recognition technologies.** Instead, we focused our efforts on alternative solutions for security that prioritize customers without compromising ethics.

Impact: By steering away from facial recognition, we have focused on developing alternative AI technologies that empower customers while remaining dedicated to responsible innovation.

CASE STUDY 2

Transparency in Human-AI Interaction

Salesforce Research: Salesforce **research** indicates that AI is more trusted and desirable when an empowered human leads the partnership with AI. "Human at the Helm" is Salesforce's version of "Human in the Loop," an approach designed to support businesses and their employees in thoughtfully steering, reviewing, and acting upon AI-generated content to advance safety, security, and trustworthiness.

The Importance of a Human at the Helm: With advancements in AI and the release of Agentforce Agents, this approach is more critical than ever. As end users interact with autonomous agents that understand and generate natural language, process and analyze large amounts of information, and take complex actions to serve the user, it is imperative that users are aware they are interacting with AI.

Policy Requirement: To ensure transparency and trust, our **AI Acceptable Use Policy (AI AUP)** mandates that customers disclose to their users when they are interacting with an AI feature, including Agentforce Agents.

Responsible AI Agent Product Design and Development

Trusted Agentic AI

To complement our ethical use policies, we also need responsibly designed and developed products. As the apex of Salesforce's AI evolution, **Agentforce** combines advanced agents, contextual understanding, and real-time learning to deliver unparalleled performance.

Salesforce's **guidelines for responsible agentic AI** provide a comprehensive framework to ensure the ethical development, deployment, and use of autonomous AI agents. They reflect Salesforce's commitment to creating AI systems that prioritize trust and align with ethical values while enabling organizations to leverage cutting-edge AI technologies effectively. Let's take a look at what this looks like in practice in our AI products:

Accuracy is a cornerstone of the guidelines, achieved through features like topic classification and clear scoping of AI agents' responsibilities. These measures work to ensure that AI agents operate only within their intended domains, minimizing errors and reducing the likelihood of "hallucinations" or incorrect outputs.

Alongside accuracy, ***safety*** is enhanced through the Einstein Trust Layer and agent-specific guardrails, which integrate safeguards like audit trails, toxicity detection, on- and off-topic classifiers, and data masking*. These mechanisms protect both the integrity of AI outputs and the privacy of sensitive information.

Transparency is central to Salesforce's principle of ***honesty***, with a focus on ensuring that users understand when and how AI is involved in interactions. Clear disclosures in AI-generated content, combined with customizable prompts and notifications, build user confidence and mitigate potential misunderstandings.

The guidelines also emphasize ***empowerment***, encouraging human oversight and collaboration with AI agents. By enabling seamless transitions between AI and human operators, these systems ensure that humans remain in control of critical decisions. Accessibility is a foundational element of this effort, promoting the empowerment of all individuals, including people with disabilities, by enhancing independence, productivity, and opportunities.

Finally, the principle of ***sustainability*** reflects Salesforce's dedication to minimizing the environmental impact of AI technologies. This is achieved by promoting the development of energy-efficient models and processes.

Collectively, these guidelines position Salesforce's agentic AI as not only a powerful tool for business innovation but also a responsible, ethical partner in advancing societal goals and maintaining user trust. By embedding these principles into its AI solutions, Salesforce aims to foster a trusted and effective AI ecosystem for the future.

*Data masking for LLMs is disabled for agents. See Data Masking and Agents. For embedded generative AI features, such as Einstein Service Replies, Einstein Work Summaries data masking is available, and you can configure it in Einstein Trust Layer setup.

Operationalizing Responsibility: Trusted Agentic AI Features

Salesforce's trusted AI features represent a significant effort to translate abstract principles of responsible AI into tangible tools and systems that ensure safety, accuracy, and transparency. By prioritizing trust as a foundational element, these features empower organizations to use AI confidently and responsibly while addressing the broader societal and ethical challenges associated with the technology.

Trust Patterns for Agents

When considering agents and their behaviors, new risks emerge. These include:

- User misuse, where AI tools might be exploited or misunderstood, leading to unethical outcomes like misinformation or privacy violations.
- Unintended agentic behavior where autonomous AI actions produce harmful outcomes, such as biased content or erroneous decisions.
- Loss of human controls, a scenario where increasing autonomy in AI agents diminishes human oversight.
- Automation bias, where users overly trust AI outputs without critical analysis, risks amplifying errors and ethical concerns.

To mitigate these risks, Salesforce has standardized guardrails implemented across our AI products, designed to improve safety, accuracy, and trust while empowering human users. We call these **trust patterns**.

These guardrails help AI systems operate within predefined boundaries and maintain transparency. For example, platform guardrails establish consistent operational protocols across all AI systems, while customer-specific guardrails are tailored to end-user interactions and aligned with industry-specific norms. These trust patterns include:

Mindful Friction: Mindful friction is a design pattern in responsible AI that intentionally incorporates checkpoints or pauses within AI workflows to encourage thoughtful decision-making and reduce the risk of unintended consequences. Unlike traditional automation, which prioritizes speed and efficiency, mindful friction encourages critical actions involving AI to be reviewed by a human who may take additional validation steps. This pattern is particularly useful in high-stakes environments, such as financial services or healthcare, where the cost of errors can be significant. By introducing moments of deliberation, users are empowered to confirm or adjust AI outputs before they are executed.

The implementation of mindful friction can take various forms, such as requiring explicit approvals for sensitive tasks, presenting alternative options, or flagging outputs that deviate from expected norms. These mechanisms enhance user accountability and reduce reliance on AI as a sole decision-maker, fostering a collaborative dynamic between humans and machines. By balancing automation with human oversight, mindful friction supports responsible AI use and aligns with ethical principles of safety and accountability.

Transparency and Notification: Transparency and notification are critical design patterns for fostering trust in AI systems. This approach facilitates users' understanding: that they are informed whenever they interact with an AI system, comprehend the role of AI in generating outputs, and are aware of its capabilities and limitations. Notifications and disclosures, such as those embedded in Salesforce's AI agents, clarify whether communications or actions are AI-generated. This transparency builds user confidence by eliminating ambiguity, allowing people to make informed decisions about how they engage with the technology.

Notifications also play a key role in educating users about AI's impact and boundaries. For example, a disclaimer in an email generated by an AI system might explain that the content was auto-generated while highlighting that it adheres to ethical guidelines. Transparent design not only fosters accountability but also protects organizations from potential misuse or misunderstandings of AI applications. By making AI systems more comprehensible, this pattern bridges the gap between sophisticated AI capabilities and everyday user interactions.

Responsible AI Product Design and Development

Citations: Citations link AI-generated responses to their original sources, allowing users to verify the source data's validity, identify potential inaccuracies or hallucinations in the AI's output, and increase confidence in using AI tools. We have implemented their inclusion via Einstein Search, which retrieves relevant knowledge data and includes it in prompts sent to the LLM. While our features present citations in various ways, they consistently connect to the relevant sources used by the LLM. This supports our Ethical Use principle of Accuracy, which allows the user to validate the AI's responses.

Prompt Design and Guardrails: Prompt design is a foundational element of responsible AI that defines how AI systems interpret and respond to user inputs. Well-crafted prompts guide AI systems to deliver accurate, contextually relevant, and ethically sound outputs. This design pattern involves structuring prompts with clear instructions, constraints, and context, helping to ensure the AI understands its tasks and operates within its intended scope. In Salesforce's ecosystem, prompt design is used to manage topics and prevent AI from attempting to address questions or tasks outside its predefined capabilities (detecting on and off-topic inputs accurately and only responding to those that are in-scope of the agent's job to be done), reducing errors and hallucinations.

Effective prompt design also incorporates safeguards against unintended behavior, such as prompt injection or misuse. For example, an AI service agent might have a topic specifically dedicated to identifying proprietary or sensitive information and redirecting such queries to human review. Additionally, prompts can be layered with ethical guidelines, steering outputs so that they align with organizational values and compliance standards. This deliberate structuring of prompts enhances both the reliability and accountability of AI systems, making prompt design an essential component of responsible AI development.

As AI evolves, the focus on trust and intentional design remains a priority. Salesforce's commitment to transparency, ethical practices, and rigorous testing lays the groundwork for an autonomous future where humans and AI work together seamlessly. Through continuous innovation and the refinement of trust patterns, AI systems like Agentforce offer a blueprint for responsible AI, empowering organizations while safeguarding users and stakeholders..

As AI evolves, the focus on trust and intentional design remains a priority. Salesforce's commitment to transparency, ethical practices, and rigorous testing lays the groundwork for an autonomous future where humans and AI work together seamlessly.

Agents Built On Top of a Trusted AI Platform

AI Security

Security is a key component to building a foundation of trust in technology. The age of AI has ushered in a new wave of security concerns that not only threaten the potential exploitation of sensitive data, but also the overall integrity and trust of the technology. Some of the top risks include:

- **Prompt injections:** Bad actors can manipulate an LLM through malicious insertions within prompts and cause the LLM to act as a "confused deputy" for the attacker. Safeguarding against these threats involves a two-pronged strategy – using machine learning defense strategies to intelligently detect and prevent malicious insertions, and using heuristic, learning-based strategies to safeguard against potential threats to prompts, such as deny list-based filtering and instruction defense

- **Data poisoning:** Attackers can change how an AI responds by injecting malicious or misleading content into the training data or documents used in Retrieval-Augmented Generation (RAG). Companies can protect against this by checking that training data or RAG documents do not contain poisoned information, such as malicious code payloads, which could compromise the model's security and effectiveness, or lead to privacy violations and other security breaches. Additionally, companies should only ground their models in content that they own, control, and verify.

- **Supply chain vulnerabilities:** Vulnerabilities can affect the entire application lifestyle, including traditional third-party libraries/packages, docker containers, base images, and service suppliers. Organizations can guard against these by ensuring that every part of the lifestyle meets the company's established security standards. And, they must ensure all components pass the company's internal security review process before they are incorporated into products.

- **Safe training grounds:** Companies should hold the training environments – controlled settings where AI systems can learn and improve their capabilities – to the same security standards as the data environment itself. This is especially important as companies increasingly view training environments as a development environment and treat them with less security.

In addition to maintaining the highest levels of security for our own AI products – including contributing to our Einstein Trust Layer, which you'll read about next – our security team has published best practices for how organizations can protect themselves from these security risks as well.

Read more in the Mitigating LLM Risks Across Saleforce's Gen AI Frontiers whitepaper.

Agents Built On Top of a Trusted AI Platform

The Einstein Trust Layer

At the core of Salesforce's AI strategy lies its robust AI platform, designed with a foundational emphasis on trust, transparency, and responsible use. This platform incorporates several key features that ensure the safe and effective deployment of AI across industries. **The Einstein Trust Layer** is a comprehensive framework within Salesforce's AI ecosystem, designed to uphold data privacy, security, and ethical standards while enhancing the effectiveness of AI applications. Its core functionalities include:

1. **Secure Data Handling:** The Trust Layer employs secure data retrieval and dynamic grounding to provide large language models (LLMs) with relevant business context without compromising data security. It maintains strict permissions and access controls, ensuring that sensitive information is protected throughout the AI processing pipeline.

2. **Data Masking* and Zero Retention:** To safeguard personal identifiable information (PII) and payment card industry (PCI) data, the Trust Layer implements data masking* techniques before transmitting prompts to third-party LLMs. Additionally, it enforces a zero data retention policy, ensuring that neither prompts nor outputs are stored after processing, thereby maintaining customer control over their data.

3. **Ethics by Design:** Salesforce integrates ethical considerations into the development and deployment of AI solutions through the Trust Layer. By adhering to an AI Acceptable Use Policy, the framework prohibits the generation of individualized medical, legal, or financial advice, ensuring that AI applications align with humane and ethical standards.

4. **Audit Trail:** Salesforce's **Audit Trail** is designed to provide transparency and accountability in AI operations. It allows admin users to track the actions and decisions of AI systems, ensuring they align with business goals and ethical standards. By documenting the entire lifecycle of an AI interaction, the Audit Trail provides a detailed log of what actions the AI agent performed, why those actions were taken, and what data or inputs influenced those decisions. This feature is particularly valuable in industries with strict compliance requirements, as it enables businesses to demonstrate how AI operates within predefined boundaries.

   The capabilities of the Audit Trail extend beyond simple logging. It empowers users to identify and address potential issues in real time, such as unexpected AI behaviors or errors. By offering visibility into the decision-making process, the Audit Trail enhances trust in AI systems and facilitates troubleshooting and optimization. Organizations can use these insights to refine AI models, improve customer experiences, and ensure alignment with both internal policies and external regulations. This robust functionality not only reinforces Salesforce's commitment to ethical AI but also provides businesses with the tools to maintain control over their AI implementations.

Continued...

*Data masking for LLMs is disabled for agents. See Data Masking and Agents. For embedded generative AI features, such as Einstein Service Replies, Einstein Work Summaries data masking is available, and you can configure it in Einstein Trust Layer setup.

5 **Real-Time Toxicity Detection: Safeguarding Interactions** AI systems are increasingly used in environments where interactions between humans and technology occur at scale, such as customer support, content generation, and social media moderation. To maintain a positive and respectful digital environment, Salesforce has developed robust **real-time toxicity detection** systems.

- **Advanced Detection Algorithms:** Leveraging cutting-edge machine learning techniques, these systems identify harmful or inappropriate content as it is generated or encountered. Detection extends across a range of toxicity types, including:

  Hate Speech: *Language intended to demean, degrade, or incite violence against individuals or groups based on attributes such as race, ethnicity, religion, gender, sexual orientation, disability, or nationality.*

  Violence, Harassment and Abuse: *Language that promotes, encourages, or depicts physical harm, threats, harassment, or sustained intimidation toward individuals or groups.*

  Sexual Content: *Language or content depicting, suggesting, or describing explicit sexual activity, nudity, or sexually suggestive themes that are inappropriate or non-consensual.*

  Self Harm: *Content that promotes, depicts, or encourages self-injury, suicidal thoughts, or harmful behaviors to oneself.*

  Drug Use: *Content that promotes, encourages, or glamorizes the use, distribution, or manufacture of illegal or controlled substances.*

  Profanity: *The use of vulgar, offensive, or obscene language, including swear words or expletives.*

- **Context Awareness:** These systems are designed to account for linguistic nuances, context, and cultural variations, reducing false positives and ensuring fairness.

- **User Controls:** Organizations can customize toxicity detection thresholds and policies, aligning AI moderation with their unique values and priorities.

Responsible AI and Tech: Trusted AI Reviews

Embedding trust into Salesforce's models, products, and features requires close collaboration with Tech and Product teams. At the heart of this effort is the Trusted AI Review process, led by **Responsible AI and Tech (RAIT)** product managers within the Office of Ethical and Humane Use. This process enables potential risks to be identified early, mitigated effectively, and tracked transparently.

How it Works

The review process begins when product teams submit a request, answering targeted questions designed to surface potential risks. These responses are scored using a due diligence framework, prioritizing high-risk products for further review. RAIT product managers then initiate a kickoff meeting and create a dedicated Slack channel to facilitate collaboration.

During the review, RAIT product managers work closely with product teams to understand the product's use cases, tech stack, and intended audience. They conduct a risk assessment to identify and categorize potential risk scenarios under **sociotechnical harm subtypes**. Each risk is evaluated both with and without mitigations, ensuring a clear understanding of its impact

Mitigation and Validation

The RAIT team works with product teams and technical experts to develop and assess mitigations for identified risks. When additional testing is needed, they collaborate with the Testing, Evaluation, and Alignment (TEA) team to validate mitigations. This process leverages both qualitative judgment and defined thresholds to ensure consistent, repeatable assessments. All review details are documented in a standardized Trusted AI assessment template and tracked using Salesforce's internal tools. Examples below in our case studies.

Outcomes

One of the key indicators of a mature responsible AI practice is how ethical risks and recommendations are operationalized within product development. At Salesforce, we treat these concerns with the same rigor as security or technical issues by filing them as "bugs" in our engineering project management system. This approach allows for ethical considerations to be actively tracked, assigned to specific owners, and resolved with clear accountability. Each issue is given a severity score that comes with a corresponding SLA (service level agreement, or agreed upon timeframe for resolution), and–like any other critical bug–has the potential to block or delay a launch if unresolved. By embedding ethical risk mitigation into our standard development workflows, we reinforce our commitment to building AI that is both innovative and responsible.

The Trusted AI Review process provides teams with a clear view of risks, mitigations, and next steps. It maintains transparency across stakeholders, aligns cross-functional teams, and ensures that AI products are developed responsibly. By embedding this process into product development, Salesforce amplifies AI's benefits, minimizes potential harms, and upholds ethical standards

CASE STUDY 1

Product Ethics Review for Agentforce Sales Development Representative (ASDR) Feature

The **Agentforce Sales Development Representative (ASDR)** feature automates outreach emails to unqualified leads and assists sales teams in qualifying these leads through bulk email campaigns. The system can respond to common inquiries, schedule calls, and redirect complex questions to human sales agents (Account Executives). Currently, the feature supports email communication, with plans to expand to web chat, SMS, and other digital channels.

Potential Risks Identified and Mitigations

1 Risk of Toxic or Unprofessional Responses

Risk: The AI agent could respond inappropriately if a lead uses profanity, offensive language, or toxic content.

Mitigation:

- The system is designed to remain professional and automatically opt-out leads who use inappropriate language.
- Validated scenarios show the agent de-escalates by removing the lead from further communications without retaliatory responses.

2 Bias and Sensitive Content Handling

Risk: The AI agent may face culturally sensitive or biased queries (e.g., regarding religious attire).

Mitigation:

- The system uses Retrieval-Augmented Generation (RAG) to provide accurate, context-aware answers.
- If the AI cannot confidently respond, it escalates the query to a human sales agent to prevent inappropriate or biased replies.

3 Data Security and Privacy

Risk: The AI agent could potentially leak sensitive CRM data or provide unauthorized information about other leads.

Mitigation:

- The AI respects user permissions and data access policies set within the product, so that it does not share data beyond a user's access level.
- The security team validated that the topic classifier and reasoning engine prevent unauthorized data exposure.

4 Autonomous Nature and Oversight

Risk: The automated nature of the email exchange could result in the AI sending emails that a sales manager may not have reviewed or approved.

Mitigation:

- Emails are batched and scheduled with a delay by default, allowing sales managers to review emails before they are sent.
- The system is designed to escalate to a human for review of any emails addressing subjects outside the agent's preset topics or not covered in the data library (RAG). That person can then handle the escalation appropriately.
- The current iteration faces challenges in providing dynamic and scalable oversight, as AI-generated communications are stored in individual lead records, making it difficult to maintain a holistic view.

  Dashboards and oversight tools are under development to address these limitations and enhance monitoring capabilities.

Continued...

Product Ethics Review for Agentforce Sales Development Representative (ASDR) Feature (continued)

5 Prompt Injection and System Abuse

Risk: Malicious users could attempt prompt injection attacks to manipulate the AI agent's behavior or extract unauthorized data.

Mitigation:

- Testing and security validations were conducted to ensure the system deflects off-topic or malicious prompts.
- The reasoning engine and topic classifier help identify and block such requests.

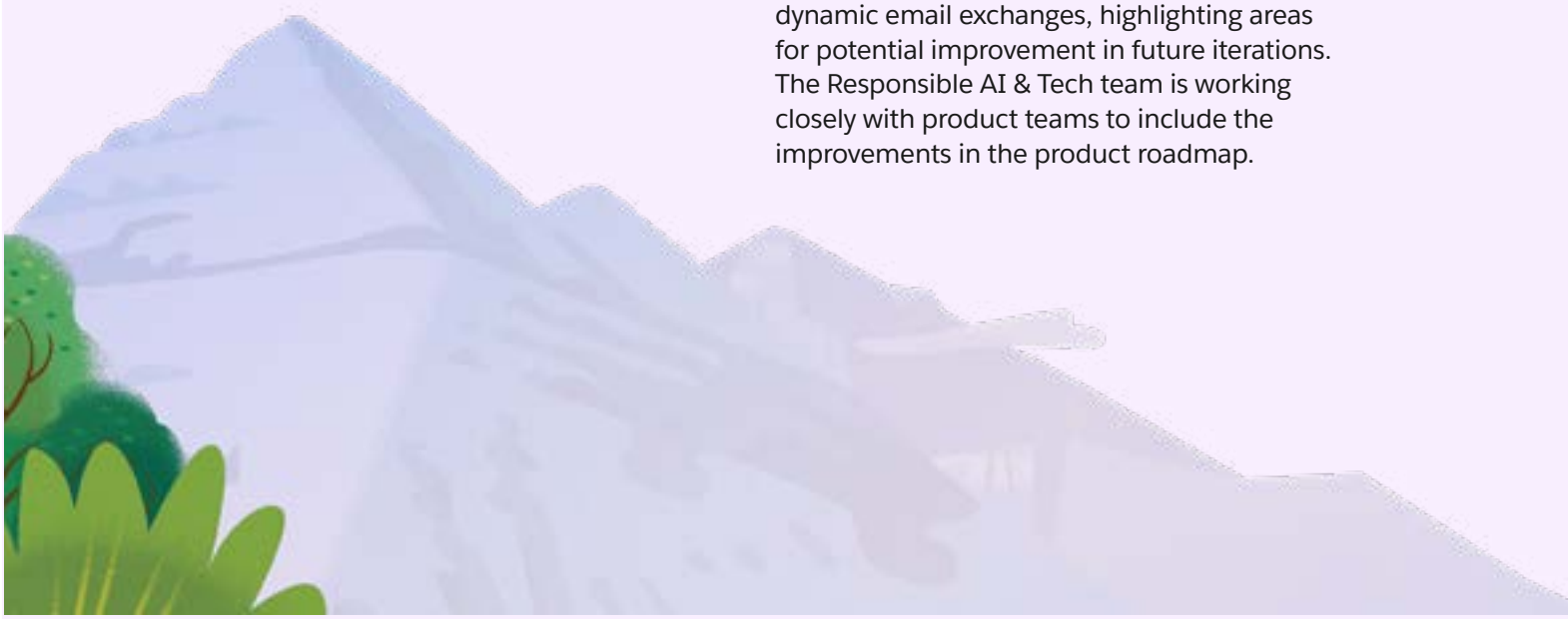
Governance and Transparency Measures

Clear Disclosure: The AI agent is explicitly identified as an AI in all communications to manage expectations and avoid misleading recipients.

Opt-Out Mechanisms: Leads can opt out of further communication either by stating their preference or clicking a link embedded in the email.

Audit Trail: All email interactions are stored in the lead's record within the CRM, ensuring transparency and accountability.

The ethics review of the Agentforce ASDR feature identified potential risks related to toxic responses, data privacy, bias, and oversight in automated communication. Mitigations, such as professional response defaults, data access controls, and human escalation pathways, help address these concerns. However, challenges remain in providing real-time oversight for dynamic email exchanges, highlighting areas for potential improvement in future iterations. The Responsible AI & Tech team is working closely with product teams to include the improvements in the product roadmap.



CASE STUDY 2

Product Ethics Review for Education Cloud Student Summarization Feature

The Student Summarization feature in Education Cloud is an AI-powered action designed to help Student Advisors (SAs) quickly prepare for student meetings. The feature generates a summary of a student's profile, including previous interactions, pulse check responses, intake assessments, and alerts. The goal is to give SAs an efficient way to get a snapshot of each student, alleviating time pressures during back-to-back appointments.

Potential Risks Identified and Mitigations

1 Bias and Stereotyping Risks

Risk: Summaries may perpetuate or amplify biases present in advisor notes, especially regarding demographic information or subjective adjectives.

Mitigation:

- Ethical containment policy instructions are included in the prompt, explicitly instructing the model to avoid demographic-based assumptions and to rely only on factual data.
- Bias testing was conducted to validate that the model does not reinforce stereotypes.
- Further testing with the ethical testing team was suggested to identify potential biases in outputs.

2 Hallucination and Inaccurate Summaries

Risk: The model might generate incorrect or misleading information, leading to opportunity loss for students if advisors rely on faulty summaries.

Mitigation:

- Summaries are grounded in data from student profiles to reduce hallucinations.
- Clear prompt instructions direct the model to avoid generating content it cannot support with data.
- Advisors retain the ability to edit and review summaries before saving them.

3 Opportunity and Service Benefit Loss

Risk: If the model generates incomplete or inaccurate summaries, advisors may need to spend additional time reviewing student records manually.

Mitigation:

- Summaries are concise (250-300 words) and designed to highlight active issues, making them easier to review quickly.
- The batching of summaries allows advisors to maintain oversight.

The **Student Summarization** feature addresses a significant pain point for Student Advisors who manage hundreds of student cases with limited preparation time.

The feature was approved to move forward with the recommended mitigations in place to protect against the identified risks of stereotyping students and model hallucination.

AI Testing, Evaluation, and Assessment

Model Benchmarking

In the rapidly evolving landscape of artificial intelligence, ensuring the trust and safety of AI products is paramount.Trust and safety (T&S) benchmarking is a critical aspect of our evaluation process. By evaluating our AI models against T&S metrics (e.g., bias, privacy, truthfulness, robustness), we can ensure that they perform at the highest level. When a model scores below a certain range on one or metric, we use adversarial testing to better understand how that manifests in practice. For example, the exact toxicity of a score doesn't tell you what kind of toxic content an LLM might generate. However, once you know that toxicity may be an issue, you can focus your adversarial testing specifically on that issue rather than testing for all types of risks.

We took our benchmarking work one step further and published the first ever **LLM benchmark for CRM** to share critical metrics essential for understanding how well an AI system operates and open up our learnings to the public. The T&S measures in the benchmark evaluates an LLM's capability to shield sensitive customer data, adhere to data privacy regulations, secure information, and refrain from bias and toxicity for CRM use cases. By assessing the reliability of LLMs for CRM, this benchmark gives organizations a sense of transparency regarding trust and safety.

Ethical Red Teaming for Trust

At Salesforce, our Responsible AI & Technology team implements **red teaming** practices to improve the safety of our AI products.

Ethical red teaming involves rigorous ethical product testing, testing adversarial scenarios, and conducting vulnerability analysis to uncover weaknesses in our AI systems before they reach customers. This tests for malicious use, or intentional integrity attacks (things that are relatively well known today, like prompt injection, or jailbreaks), as well as benign misuse (unintentionally eliciting biased, inaccurate or harmful results by a well-intended user). As described above, based on the results of our benchmarking, we perform robust AI red teaming for toxicity, bias, and security to make sure that if any malicious use or benign misuse occurs our systems are safe.

There are two main ways to go about red teaming – manual and automated – both of which are employed at Salesforce.

Manual Testing

Manual testing leverages the creativity, experience, and specialized knowledge of human testers who think like adversaries, using their expertise to craft complex and sophisticated attack strategies that automated systems might overlook. Human testers can also better understand the nuances and context of the systems they are testing and they can adapt their approach based on the specific environment, target, and goals, making their attacks more realistic and tailored.

Examples include:

Hackathons: A large group of individuals with an adversarial mindset are brought together (virtually or in person) for a specified period of time to attack your model. For example, Salesforce's **XGen Hackathon** had teams compete to identify vulnerabilities in our next-generation text generation models.

Bug bounty: These are usually conducted asynchronously and can be limited to a period of time or be permanently open for anyone to participate in. Individuals are incentivized to find vulnerabilities and report them in order to receive an award. These are excellent once a product is launched to catch new harms that weren't discovered during pre-launch. We incentivize our employees to identify and report vulnerabilities through our **Bug Bounty Program** and host **ethical bug bounties**.

Automated testing

Automated testing iis used as an enhancement, not replacement, of human-driven testing and evaluation. This type of testing involves the use of scripts, algorithms, and software tools to simulate a vast number of attacks or adversarial scenarios in a short period, systematically exploring the risk surface of the system.

One approach we've been taking to automate some of our tests is called "fuzzing," where we generate randomized test cases based on successful human attacks from manual testing (confirmed by to have been successful either in our manual testing, or through other publicly known attacks), deliver these test cases to the target model and collects outputs, and then assess whether each test case passed or failed.

Leveraging Our Ecosystem for Testing

Engaging external experts

In addition to all the work we've done internally, we have also engaged experts to perform penetration tests (through our Security Team's **Bug Bounty** program) and other creative attacks (in line with our **White House AI Voluntary Commitments**, we recently chose to outsource testing of two of our Einstein for Developers (E4D) product and our research multimodal model, PixelPlayground). Leveraging third parties can be helpful because they may approach the product and model in a completely different way than you would, offering a broader range of risks to mitigate. External experts adversarially attacked products, focusing on making the product generate biased or toxic code, while also providing unstructured attacks. We encourage others to similarly partner with security and AI subject matter experts for realistic end-to-end adversarial simulations. We will describe more about our work with external experts in a subsequent blog.

Employee Trust Testing

With the launch of Agentforce for Service, we faced a challenge: how to rigorously test for bias in a system that takes semi-autonomous action. While traditional testing methods formed the foundation of our product testing strategy, we knew that a different kind of testing was required to surface the subtle forms of bias that diverse users experience. To meet this challenge, we tapped a diverse population of employees from across Salesforce's global workforce to evaluate the trustworthiness of prompt responses. Participants engaged with the AI in simulated real-world scenarios, creating detailed personas to represent diverse user experiences and using those personas to explore interactions that probed for biases, inconsistencies, and gaps in cultural sensitivity. We called this **Employee Trust Testing**.

In other words, Trust Testing leverages diverse perspectives to understand how AI systems can maintain user trust even when outputs can't be predicted. Trust Testing expands upon our existing **ethical product testing efforts** which prioritize diverse perspectives to uncover unintentional biases in the customer experience. This serves as a cornerstone in our trust architecture, rigorously testing and identifying potential biases and vulnerabilities to ensure our AI systems are not only robust but also align with the Salesforce's standards of ethical responsibility.

As our AI continues to advance, the Testing, Evaluation, and Assessment team is dedicated to ensuring the trust and safety of our AI products. By fostering a culture of continuous improvement and innovation, we are committed to delivering AI solutions that our customers can trust.

Trusted Internal Use of AI and Agents

Our people are at the heart of Salesforce's journey toward responsible, impactful AI. We invest in fostering an environment where employees can thrive, learn, and experiment with AI ethically and productively. Below are some examples of how Salesforce is committed to employee safety, skilling, and career growth as related to AI.

AI Learning Day

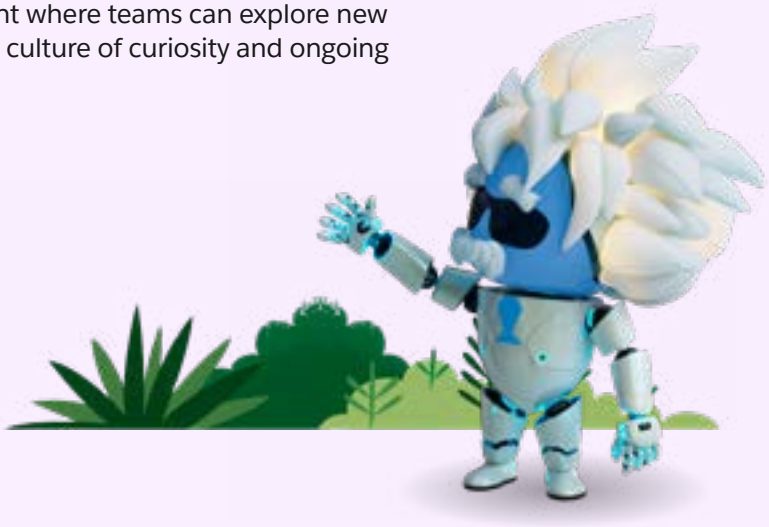
This all-company **event** represented a milestone in employee engagement with AI, equipping Salesforce teams with insights on our latest AI tools, including Agentforce. Tailored tracks, expert speakers, and both in-person and virtual watch parties provided employees across the globe the resources to deepen their understanding of AI's role at Salesforce.

Career Connect

Our new AI-powered internal talent marketplace, **Career Connect**, helps employees develop their careers by matching them with roles, assignments, and training that align with their skills and aspirations. In our pilot, Career Connect empowered 1,200 employees to explore new opportunities, with a remarkable 90% creating skill profiles and engaging with the platform. This tool enhances internal mobility and addresses skill gaps, preparing us to thrive in an AI-driven future.

Safe Sandbox for Experimentation

To promote safe innovation, we provide an environment where teams can explore new AI technologies without risk to live systems, fostering a culture of curiosity and ongoing improvement.



AI Council

The AI Council is a multidisciplinary team of experts dedicated to facilitating the responsible and swift deployment of AI tools and applications within Salesforce. Utilizing a strategic framework, the council enhances project outcomes and innovation by leveraging the guidance of teams from Legal, Ethics, Technology, and other subject matter specialists.

Empowering Employees to Use Our AI Tools

At Salesforce, we use our own AI-driven solutions to enhance productivity, from Case Summaries to Auto Replies and Voice Summaries, which streamline tasks and elevate the employee experience. Additionally, AI has helped create vision statements in our annual employee performance plans (V2MOMs), and we continue evaluating other tools, like Einstein in Slack and Basecamp, that make information more accessible and enhance our workflows.

AI and Accessibility

At Salesforce, we're committed to making AI work for everyone. We strive to develop AI products that prioritize accessibility, enabling individuals with diverse abilities to harness the full potential of technology. By embedding accessibility into our AI, we can expand equitable access to all users and empower those with disabilities to fully participate in the digital landscape.

How We're Using AI to Create More Accessible Products

Salesforce is enhancing accessibility by using AI to address specific needs across our products. For example, our **Resize/Reflow** initiative leverages AI to identify features that need to adapt to different screen sizes without losing context or function, and will eventually make interfaces more user-friendly, particularly for individuals who are Blind or low vision.

In addition, our **Self-Service for Engineering** initiative supports product teams in addressing accessibility requirements proactively. Engineers and developers can now access a repository of accessibility guidelines and answers to common questions proactively without waiting for a response from an accessibility engineer, ensuring accessibility considerations are integrated into product design and development from the start. This resource provides both foundational guidance and real-time support, making inclusive design more efficient and comprehensive across teams.

CASE STUDY 1

Building Accessibility into Agentforce

Salesforce is enhancing accessibility by using AI to address specific needs across our products. Through our Integrative Support Program, we guided our Agentforce development team in embedding accessibility at every stage, from initial design to final testing.

The program emphasized comprehensive accessibility support, including conducting design and code reviews, performing bug triage and validation, and engaging in proactive training sessions.

Before implementing these changes, Agentforce had limited accessibility features–key functionalities could only be accessed with a mouse, and crucial screen reader information was missing. After embedding accessibility support, Agentforce now supports keyboard navigation, ensuring users can interact without relying on a mouse. We also enhanced screen reader compatibility so that all interface elements are clearly described, creating a more inclusive user experience. The improvements to Agentforce illustrate Salesforce's commitment to ongoing accessibility, proving that accessibility is an evolving process that strengthens our products and benefits all users.

How We're Making AI More Accessible

Our commitment to accessibility also extends to making AI tools themselves accessible and easy to use for all users. This means designing features and functionalities that meet the diverse needs of people who rely on assistive technology and inclusive design. For example, **Alt Text Generation** is a feature in development aimed at providing visually impaired users with AI-generated descriptions of images, enhancing inclusivity across Salesforce products with automatically generated, accurate alt text.

Internally, we also focused on employee experiences to guide the inclusive design of AI tools. Our **Slack Neurodivergent SteerCo** brought together neurodivergent team members to shape Slack's accessibility features, ensuring that the platform evolves to meet the needs of all types of users. This community-driven approach fosters a supportive and inclusive work environment and was a test to ensure that we can scale and create opportunities for employee engagement in product development.

CASE STUDY 2

Accessible Citations

The **Accessible Citations Program** represents an innovative approach in our commitment to inclusive design, setting a new standard for accessible AI-generated content. Unlike our usual design process, this program explores experimental methods to create more readable and structured AI-generated citations, making them accessible for users who rely on screen readers and assistive tools. By conducting this pilot, we aim to demonstrate how inclusive design can drive innovation and inform new ways of working at Salesforce, especially in rapidly evolving AI landscapes.

This initiative goes beyond product development; it serves as a model for rapid research and validation processes that can guide other designers and internal teams as they work to meet the growing demand for accessible solutions. We hope to establish a replicable process that empowers teams to incorporate accessibility from the start, showing that inclusive design benefits everyone. By embedding accessibility in early development stages, we're fostering a culture of inclusive innovation that aims to set industry standards for AI outputs.

Our work in AI and accessibility underscores our belief that responsible technology benefits everyone. By embedding accessibility into our development processes, we're building a future where AI solutions are accessible, inclusive, and empowering.

AI for Impact

Salesforce is committed to driving positive change for our employees, the environment, and society. Our focus on Employee Success, Sustainability, and Equality reflects our core values and our dedication to responsibly innovating with AI.

Philanthropy

AI for Impact Accelerator: The Salesforce Accelerator – AI for Impact is a philanthropic initiative to help purpose-driven organizations gain equitable access to trusted generative AI technologies. The accelerator provides flexible funding, pro-bono expertise, and technology to nonprofits, empowering them to accelerate generative AI-based solutions to the world's most pressing challenges. Since 2023, the AI for Impact accelerator has helped 17 nonprofits build innovative AI solutions to advance **equity in education** and **climate mitigation, adaptation, resilience, and finance**.

Agents for Impact Accelerator: Building on the success of the AI for Impact accelerator and the company's focus on agents, we launched the **Salesforce Accelerator – Agents for Impact**, an initiative designed to help nonprofits harness agentic AI, a new layer on the Salesforce Platform that enables companies to build and deploy AI agents that can autonomously take action across any business function. This accelerator will provide technology, funding, and expertise to help nonprofits build and customize AI agents, enabling them to improve operational efficiency and scale community impact in the AI-driven future.

Philanthropic Giving: Salesforce philanthropic investments related to AI focus on championing a future where everyone benefits from AI equally. Our grantmaking in AI focuses on literacy and training, as well as tools and applications. In 2023, we gave $23 million to education to help the AI generation unlock critical skills. This funding included grants to U.S. school districts and global education nonprofits, including $6 million allocated to nonprofits focused specifically on AI skilling and literacy.

Sustainability

Our commitment to sustainability guides our product design, operations, and how we accelerate AI innovation.

Sustainable AI Blueprint: Salesforce's **sustainable AI strategy**, developed in collaboration between the AI Research, Sustainability, and Office of Ethical and Humane Use teams, focuses on optimizing models, utilizing energy-efficient hardware, and prioritizing low-carbon data centers to make our AI solutions as efficient and sustainable as possible. This strategy integrates sustainable practices at every stage, setting a high standard for responsible AI development.

Sustainable AI Policy Principles: The **Sustainable AI Policy Principles** build on Salesforce's commitment to advocate for clear and consistent science-based policies for a just and equitable global transition to a 1.5˚C future. The principles offer clear best practices for lawmakers and regulators adopting sustainable AI regulations, including how to manage and mitigate the environmental impact of AI models and ideas to spur climate innovation with policies that can incentivize and enable the environmental application of AI.

Net Zero Cloud: Net Zero Cloud AI capabilities enables companies to streamline ESG management and reporting to comply with current disclosure frameworks. By integrating AI solutions with Net Zero Cloud, customers can manage and track ESG metrics, gain accurate and timely insights, streamline reporting, create compliant reports to achieve their ESG goals.

Ecopreneurs: Ecopreneurs are purpose-driven entrepreneurs who are developing solutions to tackle some of the world's biggest challenges like climate change. Through the Salesforce Ventures Impact Fund, UpLink, and philanthropic investments, Salesforce supports ecopreneurs like PanoAI who leverage AI for the benefit of the planet.

Equality

As part of our long-standing commitment to Equality, we are dedicated to ensuring equal access to AI, and supporting the equitable development and deployment of AI technologies.

Inclusive Development of AI

Salesforce is an equal opportunity employer, which ensures we have a wide variety of perspectives to fuel innovation of our AI tools and technologies, and reduce bias across our AI systems.

Internal Reskilling

We created Career Connect, an AI-powered internal talent marketplace, to expand access to AI skills and careers for our employees.

AI for All

Beyond our four walls, we're actively working to skill up the future workforce on agentic AI to help address the skills divide and set up workers for success now – and into the future. Our AI for All initiative provides our premium AI courses and AI certifications at no cost through our free online learning platform, Trailhead, through the end of 2025.

Futureforce

We are dedicated to building the next generation of talent through our Futureforce internship programs and Futureforce Tech Launchpad, our 10-week pre-internship program that provides underrepresented students in computer science access to hands-on technical training, including how to build and use agents.

Salesforce's commitment to impactful, ethical AI goes beyond business outcomes, aiming to generate enduring value for our employees, planet, and society.

Trusted AI in Our Ecosystem

Global Responsible AI Councils

Eight Intergovernmental Partnerships: Our involvement in eight government-led groups, UNESCO's Global Business Council for the Ethics of AI, Singapore Ethical Use of AI & Data Advisory Council, Singapore AI Verify Foundation, U.S. National AI Advisory Committee, U.S. Department of Homeland Security AI Safety & Security Board*, U.S. AI Safety Institute, Oregon AI Definitions Task Force, and Washington State AI Task Force, allows us to work closely with policymakers to address emerging challenges in AI safety, privacy, and security. Through these partnerships, we help shape and uphold ethical guidelines that protect individuals and foster public trust in AI technologies.

*At time of publication, this Board is no longer active.

Industry Partnerships: Salesforce is an active member of five industry alliances, including the Data & Trust Alliance, Data Provenance Initiative, AI Alliance, WEF AI Governance Alliance, and WEF Global Future Council on Data Equity. By engaging in these partnerships, we collaborate with other technology leaders to develop standards, frameworks, and practices that promote responsible AI use across sectors, addressing data integrity, bias mitigation, and equitable AI deployment.

Global Commitments: Salesforce has signed five international commitments to date; European Union AI Pact, The White House Voluntary AI Commitments, Canada Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems, Seoul AI Business Pledge, and Trento AI Declaration. These agreements reflect our pledge to uphold high ethical standards in AI, aligning with global efforts to foster transparency, accountability, and sustainability in AI development and deployment.

Our active participation in these councils and alliances embodies our commitment to collective action for ethical AI. By working together with global leaders, we aim to contribute to a future where AI serves humanity responsibly and equitably.

Industry Information Sharing

We believe that a rising tide lifts all boats and are committed to knowledge sharing across industry, government, and civil society to advance trusted AI in society.

In the last year, we have published 20+ blog posts dedicated to the ethical and humane use of AI ranging from the top risks and related guidelines for generative AI to how we've built trust into our AI. To help our customers dive deeper, we've also created resources and guides like the National Institute of Standards and Technology (NIST), AI Risk Management Framework quick-start guide and Human at the Helm action pack. And, we published the world's first LLM Benchmark for CRM, which includes trust and safety metrics for each model.

Salesforce has also hosted and participated in a range of events for knowledge sharing across industry, with examples like our Enterprise AI Summit hosted in partnership with Eurasia Group, a public panel on Mindful Friction with Projects by IF bringing together academia, civil society, and industry, along with participation at the International Telecommunication Union's AI for Good global summit. And, we hosted and 10+ sessions focused on responsible AI principles and trust patterns built into Agentforce, our agentic AI solution, at Dreamforce, our largest conference of the year.

Conclusion

As we navigate the rapidly evolving landscape of artificial intelligence, Salesforce remains steadfast in our commitment to ethical, responsible, and human-centered AI development. This inaugural Trusted AI Impact Report represents a critical milestone in our ongoing journey to build technology that reflects our core values of Trust and Innovation.

We recognize that with great technological advancement comes profound responsibility. By embedding principles such as fairness, safety, transparency and accountability into every stage of our AI lifecycle–from research and design to deployment and monitoring–we aim to ensure that our solutions not only meet the needs of today but also anticipate the challenges of tomorrow.

This report is both a reflection of our progress and a call to action. We invite our stakeholders, partners, and the broader community to join us in shaping the future of AI–one that is innovative, inclusive, and deeply aligned with the values that unite us all. Together, we can unlock the full potential of AI, building a world where these powerful technologies benefit everyone, leaving no one behind. Salesforce will continue to lead by example, setting a benchmark for what it means to innovate responsibly in the age of AI.

This whitepaper is for informational purposes only and does not constitute legal or professional advice. No guarantees are made regarding its accuracy or applicability. Consult a qualified professional for specific guidance applicable to your situation.