Alicia Chambers
Executive Secretariat
National Institute of Standards and Technology,
100 Bureau Drive,
Mail Stop 8900,
Gaithersburg, MD 20899

Re: Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)

Comments of Salesforce, Inc.

Salesforce, Inc. ("Salesforce") appreciates the opportunity to respond to the National Institute of Standards and Technology ("NIST") request for input ("RFI") on its assignments under the Executive Order Concerning Artificial Intelligence. We welcome the efforts of the Administration to encourage the safe, secure, and trustworthy development and use of AI.

About Salesforce

Founded in 1999, Salesforce is a global leader in cloud enterprise software for customer relationship management, providing software-as-a-service and platform-as-a-service offerings to businesses, governments, and other organizations around the world. Our customers represent companies of all sizes and across all sectors. Our business model is cloud-based and low-code, allowing for faster deployment of technologies and greater agility. We help our customers connect with their customers – or employees or citizens – in a whole new way using cloud, data, and AI technologies.

Salesforce & AI

Trust is our number one value and at the center of our enterprise business model – our customers' data belongs only to them. Even prior to the current conversations on regulating responsible AI governance, we have been committed to delivering safe and trusted AI to our enterprise customers to meet their expectations of high levels of privacy, security, and accuracy. These also form part of our contractual agreements with our customers. Furthermore, as a service provider entrusted with the data of companies large and small, in multiple jurisdictions, and across many sectors, Salesforce is in a unique position to observe global trends in AI technology and identify developing areas of risk and opportunity. We not only consider the impact to our customers, but also our customers' customers.

Salesforce's AI capability, called "Einstein," is built into the Salesforce platform, and is designed to combine artificial intelligence with Salesforce's suite of enterprise services, democratizing the power of AI for every Salesforce user. Powering this innovation is Salesforce Research, which was first established in 2014 and has published 200+ research papers and registered 300+ AI patents.

Many of the use cases for both our AI and generative AI ("GAI") technologies involve creating efficiencies – for example: suggesting to a salesperson the best way to engage with a potential customer, or helping that salesperson draft an introductory email. These use cases are generally lower-risk and augment human decision-making processes.

We believe that the tremendous benefits of AI should be accessible to everyone while ensuring that those technologies remain safe and inclusive. At Salesforce, we are committed to providing our employees, customers, and partners with the tools they need to develop and use safe, accurate, and ethical AI. We are committed to a vision of ethical AI that is:

- **Responsible**: To safeguard human rights and protect the data we are entrusted with, we work with human rights experts, and educate, empower, and share our research with customers and partners to enable them to use AI responsibly. We strive to comply with the laws and values of the markets in which we operate and to adhere to the highest security and safety protocols.
- **Accountable**: Accountability to customers, partners, and society is essential. Independent feedback should be sought to continuously improve practices and policies and to mitigate harm. We seek guidance and feedback from diverse stakeholders and from our Ethical Use Advisory Council.
- **Transparent**: Our customers should be able to understand the "why" behind each AI-driven recommendation, output, and prediction so they can make informed decisions, identify unintended outcomes, and mitigate harm. We strive for model explainability and clear terminology to ensure customers are informed and in control.
- **Empowering**: AI is best utilized when paired with human ability, effectively augmenting people and enabling them to make better decisions. Accessible AI promotes growth and efficiency, and benefits society as a whole.
- **Inclusive**: AI should respect the values of all those impacted, not just of its creators. To achieve this, we test models with diverse data sets, seek to understand their impact, and build inclusive teams.

Because of the rapid evolution of the technology, along with the opportunities and challenges emerging from the use of generative AI, we have gone one step further and articulated an additional set of guidelines for the development of trusted and responsible GAI at Salesforce and beyond.

- **Accuracy**: We seek to deliver verifiable results that balance accuracy, precision, and recall in AI models by enabling customers to ground models on their own data. We are researching ways to enable customers to verify the accuracy and assess the suitability of content through things like citing the source the content was drawn from and providing relevancy scores.
- **Safety**: As with all of our AI models, we make every effort to mitigate bias, toxicity, and harmful output by conducting bias, explainability, and robustness assessments, and red teaming. We also work to protect the privacy of any personal data used for training and create guardrails to prevent harms.

- **Honesty**: When collecting data to train and evaluate our models, we confirm data provenance and ensure that we have consent to use the data. We also disclose to end users when AI content they receive is being generated by an autonomous agent, like a chatbot, rather than a human.
- **Empowerment**: There are some cases where it is best to fully automate processes, but oftentimes AI should play a supporting role – such as when human judgment is required. We seek to identify the appropriate balance to "supercharge" human capabilities and make these solutions accessible to all (e.g., generate ALT text to accompany images).
- **Sustainability**: We aim to develop right-sized models where possible to reduce our carbon footprint. Larger doesn't always mean better: Focused, highly trained models often outperform generalized, minimally trained ones.

As AI becomes ubiquitous in the modern economy, policymakers and industry leaders should work together to establish guardrails ensuring the ethical development and utilization of this powerful tool.

Salesforce is committed to building trusted, transparent, and accountable AI systems that prioritize fairness, accuracy, privacy, and positive societal impact.

1. Developing Guidelines, Standards, and Best Practices for AI Safety and Security

NIST AI RMF

NIST has proven to be one of the most effective organizations for the establishment of globally recognized standards that are comprehensive, collaborative, and dynamic. The NIST Cybersecurity Framework is recognized as a global best practice, and now with the NIST AI Risk Management Framework (RMF), the Administration is working to provide a flexible cross-sectoral guide to tackling the risks associated with AI and GAI. Salesforce has and will continue to engage with NIST's AI RMF and Playbook, which include many of the measures we have internalized as we develop our own AI technology. Finally, we appreciate the work being done by NIST to ensure the AI RMF is and remains relevant to GAI.

Roles & Responsibilities In The Value Chain

Every entity in the AI value chain has a role to play to ensure the responsible development and use of the technology. Regulation should take into account the complexity of AI ecosystems and assign risk-proportionate obligations to each party in the value chain, according to their role and the level to which they control (i) the data that goes into the AI system, and (ii) the context in which the AI is used.

Creators of AI ("AI Developers") often create general customizable AI tools, of which the intended purpose is low-risk. It is then up to the customer or user ("AI Deployer") to decide how these tools are employed. In the B2B context, it is often the customer who ultimately controls what data is submitted to the AI, how the AI is configured, when to use the AI and in which context, and, most critically, how the AI's outputs are used. Frequently, it is only the customer who knows what has been disclosed to individuals potentially impacted by those outputs, and what the risk of harm is to the individuals.

In most cases, B2B companies like Salesforce do not own or control their customers' data. As a data processor/service provider, we use data at the instruction and on behalf of our customers. Consequently, Salesforce and similarly situated companies often will not be in the best position to know or control how customers use their own data in an AI system. There should be a reasonable division of responsibilities between entities situated at different points along the AI supply chain, with B2B companies like Salesforce primarily responsible for educating and enabling our customers and partners to deploy AI safely and responsibly. Further, Salesforce has instituted mechanisms like our AI Acceptable Use Policy (AUP), which act as a guardrail to safeguard against high risk use cases.

Transparency

Transparency for enterprise AI can differ from transparency for consumer-facing AI. Since Salesforce's generative AI-enabled tools and services are designed to be tailored by our customers for use in diverse business contexts, our customers expect us to not only provide tools to better understand AI-driven recommendations they receive, but to assist in providing that same level of transparency to their own end users.

We are meeting this expectation by publishing model cards for our predictive AI models (and, soon, for our GAI models), by maintaining an AUP, and by providing model explainability measures when AI predictions or recommendations are made. Salesforce has worked toward building robust explainability measures for predictive AI, but explainability is more difficult for GAI. Nevertheless, providing and improving explainability in both predictive and generative AI remain a priority for us.

Most, if not all, good governance best practices for GAI require both Developers and Deployers to record and track the performance of their models across a variety of applications, their underlying datasets, their sources, and their training techniques – similar to how third-party components are tracked in other products.

Because the AI landscape is shifting so rapidly, we regularly evaluate and adapt these processes as new developments come to light. We expect this to be the same for most other actors in the AI ecosystem, although their differing roles along the supply chain may necessitate them taking different, but valid, approaches. At times, attempts at benchmarking by third parties have failed to capture these nuances, leading to inapt comparisons between actors leading to reputational harm. While Salesforce supports NIST's development of a neutral, standardized approach to third-party transparency benchmarking and evaluation for generative AI models, any such standards or guidance must allow for a degree of flexibility in recognition of the diversity of valid approaches to AI governance.

Audits

Auditing system usage can diagnose potential or real security issues and monitor for unexpected changes, usage trends, and abuse. In general, auditing features don't secure an

AI system by themselves. However, having established guidance around auditing can be helpful for organizations seeking another layer of trust.

Salesforce supports NIST's development of guidance and/or benchmarks for the design of test environments for AI technologies. Because risk management for generative AI is a constantly evolving process, we expect there to be an increasing need to work on fine-tuning customized models to take into account newly identified risks. A baseline set of parameters for testing environments would be particularly useful for strengthening trust among AI Developers and Deployers.

Human Feedback

Salesforce strongly supports the development of best practice guidance regarding structured mechanisms for gathering human feedback. As mentioned above, different actors in the AI ecosystem play varied roles to keep the ecosystem safe for everyone. In addition to testers, end users also have a role to play. For all the guardrails put in place by Developers and Deployers to mitigate risks and potential harms, end users are often among the first to identify potential harms, workarounds, and exploits and should be given an avenue to report them.

Testing and Red Teaming

One unique challenge Salesforce faces as a provider of enterprise AI solutions is having to anticipate potential risks or harms once our customers build their applications on top of our platforms. With this in mind, we only build products and services that we know can be developed and deployed safely, as guided by our *Trusted AI Principles*. We also conduct consequence scanning, risk analyses, and robust testing (both functional and adversarial) of particular applications and models.

Our adversarial testing - or red teaming - includes evaluation of our products' vulnerabilities to risk vectors such as misalignment with human values and preferences, the introduction of bias or toxicity, confabulation and inaccuracy, data privacy and security, prompt injection and social engineering, and more. If and when we identify issues for remediation, we work to address those vulnerabilities prior to launch, in order to align with our commitments to Safety. It's important to note that we focus our red-teaming not only on our models, but also our apps. This level of testing allows for more concrete and specific use cases to be evaluated and results in increased confidence and trust in the safety of the apps themselves. We have manually red-teamed 18 products to prevent bias, toxicity, and ensure alignment with our ethical AI commitments and plan to scale this in the coming year.

As mentioned previously, the necessary skills and backgrounds of AI red teams differ significantly at the various stages of the AI lifecycle, and this should be reflected in any further guidance published by NIST on generative AI.

One unifying factor for most, if not all AI red teams, however, is their potential exposure to some of the worst content on the Internet. In many situations, AI red teams may be on the front lines in terms of preventing outputs of CSAM or gross depictions of violence with many potential consequences to their mental health and wellbeing. When developing standards for AI red-teaming, therefore, Salesforce advocates that NIST develop, either on its own or in collaboration with relevant federal agencies such as OSHA, minimum standards of working conditions and/or other protections for AI red team members.

2. Reducing the Risk of Synthetic Content

First, Salesforce would encourage techniques around identifying sources or references to help validate the factual correctness of the content being generated, explaining the reason why the AI behaves in a certain way (explainability), and indicating the level of confidence about the factual correctness of the content being generated be included in the Secretary's report.

We caution against relying too heavily on watermarking as a primary means of detecting synthetic content generated by AI. At the time that the generative AI governance discourse began to gain momentum, watermarking was frequently cited as a viable alternative to passive detection techniques that had already been shown to not be robust to evasion.

However, recent research has demonstrated flaws with current watermarking techniques that could allow malicious actors to evade watermark detection using publicly available code. Although watermarking has several benefits and should be included in NIST's development of standards and guidance on synthetic content detection, we are concerned by an overemphasis and overreliance upon it at the exclusion of other methods.

One promising technique not discussed in the RFI is *retrieval*, which involves cross-referencing a body of text against a database of previously generated AI outputs.[1] Retrieval is particularly useful for identifying bodies of text that are paraphrased by humans after being generated by an LLM, which is one of the major challenges faced by other AI text detection techniques.

Retrieval is not perfect. It requires the model owner to front storage costs, introduces privacy concerns, and has been shown to be vulnerable to recursive paraphrasing.[2] That said, we believe retrieval warrants further study and attention by NIST.

3. Advance Responsible Global Technical Standards for AI Development

[1] Krishna Kalpesh, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. "Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval Is an Effective Defense." arXiv, October 17, 2023. https://doi.org/10.48550/arXiv.2303.13408.

[2] Sadasivan, Vinu Sankar, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. "Can AI-Generated Text Be Reliably Detected?" arXiv, June 28, 2023. https://doi.org/10.48550/arXiv.2303.11156.

Given the global nature of many actors in the AI ecosystem, there must be interoperability between international standards, schemes, and protocols. The emergence of fragmented policies and approaches to the responsible development and deployment of AI systems can raise the cost of operations and undermine consumer protections. Therefore, as policymakers think through a range of possible responses to the opportunities and challenges presented by AI applications, they should prioritize finding global consensus and interoperability in venues like the G7, the Organization for Economic Co-operation and Development ("OECD"), and the US-EU Transatlantic Technology Council.

Considerations specific to enterprise AI data management should also be taken into account. Within the Salesforce platform, we have natively built our Einstein Trust Layer specifically to allow businesses to benefit from AI without compromising their data. We allow users of our platform to limit how much of their data interacts with our AI systems and to ground their generative AI prompts on customizable specifications. Salesforce technologies can automatically detect the presence of personal data and allow users to mask such data before it interacts with LLM. Lastly, to ensure customers have control over the use of their data, our LLMs forget both the prompts and outputs once the outputs are delivered.

Data Ethics

AI consensus standards, cooperation, and coordination should be rooted in globally recognized privacy principles, such as those set forth in the OECD's *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*. Personal data should be obtained lawfully and with the informed consent of the data subject. Personal data should only be used for the purposes it was collected for and should be accurate and kept up to date. It should only be retained for as long as it takes to fulfill the stated purpose of its use and disposed of afterward. Salesforce's is committed to protecting personal data by adhering to (and encouraging other stakeholders' adherence to) the following core principles:

- Treat Sensitive Data With Heightened Care: Some data types present a more significant risk of harm than others, such as health information, race, ethnicity, political and religious belief, and sexual orientation.
- Collecting and Use Only What You Need: Limit the collection and use of personal data to only what is needed to create more personalized, useful experiences for end users. Less is more, especially when it comes to demographic, socioeconomic, behavioral, and transactional data.
- Choose Partners Carefully: In addition to being mindful and intentional about selecting and ingesting first- and third-party data, organizations should be intentional about activating that data through various platforms. When engaging with an activation partner, it is critical to monitor the chain of custody of any data provided to that partner – for example, whether and when it is deleted or reused.
- Data Deletion and Retention: Organizations should only store personal data for as long as it's required and for the originally intended purpose of that data. Clear

rationales should be established for the retention of data and clear time frames should be established for its deletion.

Data Security

Organizations should implement comprehensive technical and organizational security measures to protect personal data against unauthorized processing, accidental disclosure, loss, destruction, and alteration. We acknowledge that the security of our systems is only as robust as their most vulnerable component, hence we adopt a comprehensive approach to security, scrutinizing all elements to mitigate potential risks. Salesforce uses methods such as:

- A comprehensive AI security review process: We meticulously review all components that directly or indirectly interact with an AI technology workflow. This encompasses both our proprietary products and those of vendors, and continually adapts as existing vendors incorporate AI technologies.
- Data Segregation: We require customer data to be stored separately and securely throughout its interaction with AI technologies - from ingestion, sourcing, and training, to customer usage for inferencing, to collected feedback.
- Encryption: We utilize sophisticated encryption techniques and solutions to safeguard data during transmission and storage across every touchpoint for AI technologies.
- Access Control: We follow the principle of least privilege for every step of an AI workflow and seek alternative methods (e.g. synthetic example data) when possible to minimize potential exposure of sensitive information.
- Data Masking: We adopt a comprehensive approach to the application of AI technologies in our products and implement compensating controls to minimize the potential risk of exposure and misuse across all applications of AI technologies - be it training or inference, by employing various methods such as:
    - Pseudonymization: We substitute personally identifiable information with pseudonyms, thereby reducing the risk of data breaches and enhancing privacy.
    - Anonymization: We eliminate or modify information that identifies individuals, rendering it impossible to link data back to them.
- Data Usage Transparency: We are steadfast in our commitment to transparency in our data usage. We clearly articulate how we use data and empower individuals with control over their personal information.

Biases & Fairness

One of the key challenges in data-driven decision-making is the presence of bias, which can lead to unfair or discriminatory outcomes. Bias can be introduced at any stage of the data lifecycle, from data collection to algorithmic decision-making.

Addressing bias and promoting fairness requires a range of strategies, including:

- Diversifying data sources: One of the key ways to address bias is to ensure that data is collected from a diverse range of sources. This can help to ensure that the data is representative of the target population and that any biases that may be present in one source are balanced out by other sources.
- Improving data quality: Another key strategy for addressing bias is to improve data quality. This includes ensuring that the data is accurate, complete, and representative of the target population. It may also include identifying and correcting any errors or biases that may be present in the data.
- Conducting bias audits: Regularly reviewing data and algorithms to identify and address any biases that may be present is also an important strategy for addressing bias. This may include analyzing the data to identify any patterns or trends that may be indicative of bias and taking corrective action to address them.
- Incorporating fairness metrics: Another important strategy for promoting fairness is to incorporate fairness metrics into the design of algorithms and decision-making processes. This may include measuring the impact of certain decisions on different groups of people and taking steps to ensure that the decisions are fair and unbiased.
- Promoting transparency: Promoting transparency is another key strategy for addressing bias and promoting fairness. This may include making data and algorithms available to the public and providing explanations for how decisions are made. It may also include soliciting feedback from stakeholders and incorporating their input into decision-making processes.

Adopting these strategies helps organizations ensure their data-driven decision-making processes are fair and unbiased. To ensure that AI and machine learning are developed and deployed in a responsible and ethical manner, it's important to have ethical frameworks and guidelines in place.

Finally, data privacy standards can aid in ensuring that some of the challenges around bias and fairness are mitigated such as categorizing and inventory data, analyzing risks related to processing (including processing for AI purposes), and requiring enhanced privacy obligations as the potential risks increase. However, there is an existing tension between privacy and bias mitigation, that will require statuary change but should be acknowledged in standards. An October 2023 Stanford HAI policy brief entitled, *The Privacy-Bias Trade-Off,* acknowledged the tension saying, "adoption of data minimization in the Privacy Act of 1974 has brought many privacy benefits but stymies efforts to gather demographic data to assess disparities in program outcomes across federal agencies". While the paper was about the federal government, it can be expanded out to the whole of the AI ecosystem. While data minimization is important, the lack of demographic data hampers the ability to perform bias assessments. While we understand the NIST is compiling information for guidance, Salesforce believes it is also important to surface current obstacles. As NIST moves forward on work on standards, it would be helpful, where appropriate, to acknowledge ongoing obstacles in ensuring trusted AI.

Conclusion

Salesforce is committed to building trusted, transparent, and accountable AI systems that prioritize fairness, accuracy, privacy, and positive societal impact. We welcome the focus on AI systems, including generative AI, by legislators, regulators, and other policymakers. AI raises critical and rapidly evolving questions for society, and Salesforce is proactively engaging with governments and other stakeholders to advance responsible AI public policies. Government and industry working together with other stakeholders to establish a common approach on definitions, roles, and obligations will create more durable, robust, and interoperable AI norms.

We are encouraged by the Administration's interest and look forward to further engagement with NIST. Salesforce remains committed to the success of our customers and we view our active participation in this important national discussion as advancing that success. We would be pleased to serve as a resource to NIST as it further develops its approach on trusted AI.

###