Business and Human Rights Journal (2026), 1–6
doi:10.1017/bhj.2026.10055

Business and Human Rights Journal

DEVELOPMENTS IN THE FIELD

Leveraging Human Rights Frameworks for Agentic AI Governance

Angeline Lee and Shahla Naimi

Office of Ethical and Humane Use, Salesforce Inc, USA
Corresponding author: Angeline Lee; Email: angeline.lee@salesforce.com

Abstract

Agentic artificial intelligence (AI)—systems that reason, act, and interact with autonomy and persistence—presents qualitatively new governance challenges compared to traditional large language models. By operating across multimodal environments and pursuing goals without constant human input, agentic AI introduces both opportunities to advance human rights and risks of opaque, distributed responsibility for harms. This paper argues that the UN Guiding Principles on Business and Human Rights (UNGPs) provide a practical foundation for governing agentic AI. We reinterpret the Protect–Respect–Remedy pillars to map onto agentic capacities: reasoning requires anticipatory due diligence; acting demands safeguards and oversight; and interacting calls for transparency, remedy, and accountability. We propose shared responsibility frameworks, grounded in interoperability standards, ongoing monitoring, and redress mechanisms, as essential governance tools. Ultimately, embedding human rights frameworks into the design and deployment of agentic AI enables both risk mitigation and the active promotion of rights in complex, multi-actor ecosystems.

Keywords: Agentic AI; Human Rights; Responsible and ethical AI; Salesforce; UNGPs; Shared responsibility framework

1. Introduction

This paper explores how human rights frameworks can assist in the governance of agentic artificial intelligence (AI) systems. Agentic AI refers to systems capable of acting independently to achieve goals, representing a transition from AI as a 'thinking' tool to a 'doing' tool. To address the risks inherent in this shift, this analysis is anchored in the United Nations Guiding Principles on Business and Human Rights (UNGPs)—the authoritative global framework for defining corporate responsibility to respect human rights. This paper examines how the '*Protect, Respect and Remedy*' framework may be adapted to the emerging governance challenges posed by agentic AI.

As authors working in the Office of Ethical and Humane Use at Salesforce, we approach this topic from a dual perspective: as practitioners building these systems and as scholars engaged in the broader discourse on business and human rights. This dual role informs our view that while agentic AI is a transformative enabler of rights, it also introduces qualitatively new vectors for risk.

© Salesforce, 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-ShareAlike licence (http://creativecommons.org/licenses/by-sa/4.0), which permits re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 11:46:52, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bhj.2026.10055

2. Understanding Agentic AI and its Governance Implications

Unlike traditional generative AI systems that respond to discrete individual prompts, agentic systems operate with persistence, adaptability and goal-directed planning over extended time horizons. They function across wide-ranging multimodal environments, connecting with APIs, websites and other AI agents to complete complex tasks. This results in a complex ecosystem characterised by three core capacities: reasoning (goal setting and strategy adaptation), acting (task execution) and interacting (engaging with humans or other agents).

Risks can arise at each of these points, as defined by Plaat et al (2025).[1] For instance, an AI-powered delivery agent may autonomously prioritise servicing dense urban areas over rural communities to optimise costs, resulting in discriminatory outcomes and unequal access to essential services. Similarly, a healthcare scheduling agent might prioritise patients with routine needs over those with complex chronic conditions to maximise throughput, effectively disadvantaging vulnerable populations. In each case, the system's autonomous goal-seeking behaviour—rather than an explicit human directive—produces outcomes that may conflict with human rights norms.

These dynamics illustrate why agentic AI is not just an incremental upgrade to traditional generative AI systems but a fundamentally new ecosystem. The combination of autonomy, persistence and multi-agent interaction introduces qualitatively distinct governance challenges. By design, AI agents can set and pursue goals without constant human input, retain context over extended periods, adapt across multiple environments and depend on other agents in ways that are difficult to anticipate and control. This combination is what makes AI agents powerful, enabling systems to move from suggesting a vacation to autonomously planning and booking it end-to-end.

These same features explain why agentic AI can both promote human rights and raise new questions about responsibility, oversight and risk. For example, an AI system used in healthcare operations that autonomously triages patients and allocates resources could disadvantage marginalised groups if it relies on biased historical data or cost-based proxies for need. As it operates continuously across decisions, these disparities can scale and persist unnoticed without meaningful human oversight.

These qualities matter for human rights governance because they introduce new forms of opacity and diffuse responsibility. They challenge existing assumptions about who is responsible for causing, contributing to or being directly linked to harm. In many ways, this builds on debates around distributed agency, in which responsibility is shared across humans and machines. Agentic AI operationalises these debates, requiring abstract ethical questions to be translated into concrete governance mechanisms spanning multiple actors —human and artificial—across companies, jurisdictions and technical systems. Responsibility structures are further strained because harms may surface long after deployment, complicating causal tracing and accountability. In addition, agents may act without direct human instruction, raising difficult questions about accountability, oversight and safety guardrails.

Alternatively, this same autonomy that introduces risk is also central to agentic AI's capacity to promote rights at scale. Properly designed and governed, agentic systems can autonomously coordinate disaster relief logistics, dynamically tailor accessibility supports for individuals with disabilities and facilitate civic participation by proactively connecting people to public services and democratic processes—functions that traditional, non-agentic

1 Aske Plaat, Max van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg, 'Agentic Large Language Models, a Survey' (2025) *Journal of Artificial Intelligence Research* 84, Article 29. https://doi.org/10.1613/jair.1.18675

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 11:46:52, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bhj.2026.10055

AI systems are not equipped to perform. This dual imperative—to mitigate risks while enabling rights-promoting outcomes—underscores why agentic AI governance requires adaptation of existing frameworks. Keeping this dual goal in view helps explain why human rights frameworks, particularly the UNGPs, are well suited to the governance of agentic AI.

3. Reinterpreting the UNGPs for Agentic AI

The UNGPs centre on the Protect–Respect–Remedy framework, which assigns businesses the responsibility not only for harms they directly cause but also for those they contribute to or are linked to through their products and services. Applying this framework to agentic AI requires connecting system features to each pillar:

- **Reasoning capacity (planning and goal setting):** This aligns with the Respect pillar and necessitates anticipatory human rights due diligence (HRDD) prior to deployment, as planning functions shape the agent's scope and decision logic. States also retain a Protect duty to establish regulatory baselines for *ex ante* risk assessment.
- **Acting capacity (autonomous execution):** This implicates both the Respect and Protect pillars. It demands built-in safeguards, constraints and oversight mechanisms to prevent harmful actions from being executed without human review, particularly in high-risk sectors.
- **Interacting capacity (engagement with humans or agents):** This connects to the Remedy pillar. Where transparency gaps or unfair interactions shape user experience, accessible and effective grievance and redress mechanisms are essential.

Consider a finance agent deployed to assist with loan prequalification that systematically denies applicants based on zip code. In this scenario, multiple actors are involved: the deploying company, the platform provider and the model vendor. The UNGPs encourage upstream actors to use contractual terms, technical constraints and policy enforcement to prevent harm. This treats leverage as a governance tool rather than relying solely on liability frameworks. It may require language model providers to mandate human oversight for high-stakes decisions (like loan denials) or restrict the use of protected characteristics in specific use policies.

In many respects, these complex operating ecosystems—spanning humans, companies and AI agents—test the limits of the UNGPs. Designed around human-centred decision chains, the UNGPs must be reinterpreted when applied to systems that operate with persistence and autonomy, prompting a reassessment of what it means to 'cause', 'contribute to' or be 'directly linked to' harm. By mapping the UNGPs onto the reasoning, acting and interacting capacities of agentic AI, the framework moves beyond static, point-in-time assessments towards a more dynamic model of systemic accountability. This shift has concrete implications: it clarifies how developers, platform providers and deployers share responsibility not only for discrete outputs but for the ongoing behaviours and downstream impacts of agentic systems over time.

4. Operationalising Shared Responsibility in Practice

Causal opacity in autonomous systems often blurs the boundaries of business accountability. An adverse outcome might involve the customer, the platform provider, and/or an upstream model vendor. Agentic AI requires a shared accountability model that incorporates both technical connections and multi-actor management.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 11:46:52, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bhj.2026.10055

At Salesforce, we operationalise our human rights commitment by providing customers with tools and default safeguards that promote responsible and rights-respecting AI use. For example, our products offer privacy controls, customisable fields and filters and audit logs to support responsible deployments of the Agentforce platform. We also communicate clear usage policies through an AI Acceptable Use Policy, aligned with our responsible AI commitments and broader company values. These efforts represent an initial roadmap for how platform and model providers can begin to put human rights management into action through a mix of technical, contractual and policy tools—a continuous journey of investment, refinement and collective learning for all stakeholders.

The starting point for companies to explore similar shared responsibility models lies in two primary activities:

1. **Translate the UNGPs into actionable governance steps**—for example, through policies (e.g., recommendations, usage policies), processes (e.g., required human review, continuous monitoring) and product features (e.g., built-in guardrails).
2. **Design for rights, not just risk mitigation**—embed human rights as an architectural principle rather than as an afterthought.

Practically, companies can address the dual imperative of leveraging agentic AI's opportunities while mitigating risks by prioritising the following, as it makes sense for their offerings:

1. **Continuous monitoring:** AI agents adapt and evolve based on new data, so one-time human rights checks are insufficient. Ongoing human rights due diligence (HRDD) is essential, particularly in multi-agent environments managed by multiple companies, where harms can emerge months after deployment.
2. **Shared responsibility:** Responsibility is distributed across developers, platform providers, customers and external vendors. Under the UNGPs, 'contributing' to harm can include enabling or insufficiently constraining harmful agent behaviour, even when downstream users carry out the actions. Open standards—such as Agent-to-Agent (A2A) protocols and the Model Context Protocol (MCP)—can establish governable multi-agent ecosystems. These standards strengthen the Respect pillar by embedding safeguards in agent specifications and the Remedy pillar by ensuring every interaction is auditable.
3. **Interoperable design:** When agents summon tools or delegate tasks across vendors, metadata trails can map causation, contributions and linkage—addressing UNGPs' '*linked to*' responsibility. These protocols can provide technical visibility into shared responsibility by mapping cause (e.g., direct code/decision), contributions and linkages across the ecosystem. Embedding accountability into interoperable infrastructure allows prevention, oversight and remedy to be system defaults rather than optional add-ons.

These core steps should be operationalised by embedding human rights framework best practices directly into existing safety and responsible AI processes, including:

- **Red teaming and adversarial testing:** Beyond technical robustness, test for rights-based edge cases, especially in high-risk industries (e.g., how an agent might infringe on privacy or generate prohibited content).
- **Stakeholder consultation:** Incorporate external expertise to surface harms internal teams may not foresee. Agentic AI demands a more pluralistic form of risk anticipation.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 11:46:52, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bhj.2026.10055

- **Ongoing HRDD:** Adapt HRDD processes for unique agentic AI risks like autonomy, emergent behaviour and unpredictability. Assessments should be conducted on an ongoing basis and include scenarios of unforeseen AI actions due to poor data, dynamic environments or user prompts.
- **Use limitations and policy safeguards:** Establish acceptable use policies (AUPs) tailored to agentic behaviours, making allowed and prohibited uses to customers clear (e.g., restricting autonomous actions in high-risk or regulated industries like health or criminal justice without human oversight).
- **Explainability at the point of action:** Build systems that can log, surface and explain autonomous actions. Audit logs should include who deployed the agent, what environment it acted in and why it took a given action as a baseline for later human review.
- **Tiered controls for autonomy:** Allow customers to configure how much freedom agents have (e.g., read-only versus execution capabilities) based on their own governance maturity and provide training and resources to customers to help inform their decisions.
- **Accessible grievance mechanisms:** Update user-facing grievance mechanisms to handle harms potentially caused by agents. End users should not have to understand how a system works in order to seek and obtain remedy.

No single actor can fully respect and protect rights in agentic systems alone. Implementing the UNGPs in agentic AI requires cross-industry collaboration, open standards, shared incident response and global coordination to align legal frameworks with UNGP-based approaches. Responsibility is shared: providers set the basic rules, while customers control deployment, monitoring and ethical use. Shared risk requires shared accountability for human rights.

5. Agentic AI as an enabler of rights

While governance often emphasises risk mitigation, it is equally important to recognise agentic AI as a significant enabler of human rights. Governance should address two complementary objectives: mitigation (preventing harm) and promotion (advancing rights). Preventing harm alone does not generate positive rights outcomes, and promoting rights does not offset existing harms. Both objectives must be pursued and evaluated concurrently.

Agentic AI's autonomous, persistent action allows it to overcome logistical barriers and advance rights that are otherwise difficult to realise, such as equitable access to work, education and civic participation. Examples include:

- Matching job seekers to opportunities in real time.
- Dynamically tailoring educational content to individual students.
- Guiding citizens through complex government services.

Agentic AI can also support environmental and public-interest outcomes, such as optimising supply chains to reduce pollutants, coordinating disaster relief or managing public health resources efficiently. The ways agentic AI systems can serve as an enabler of rights are endless, but they are dependent on thoughtful design and deployment as well as broader investments in infrastructure, governance and skills growth.

In other words, agentic AI is not a silver bullet. The realisation of these benefits depends on material conditions such as affordable connectivity, as well as on thoughtful design and

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 11:46:52, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bhj.2026.10055

deployment practices, including linguistically and culturally competent system design. These benefits also require sustained governance to ensure that the same autonomy used to promote rights does not simultaneously undermine them.

For Salesforce, this dual focus aligns with broader AI and human rights efforts, such as integrating accessibility features into Agentforce workflows or deploying agents in public sector services to meet the needs of different populations. Viewing both risks and opportunities together mitigates the risk that agentic AI could reinforce inequalities and instead sets up agentic AI as an enabler of rights realisation at scale.

6. Conclusion

Nearly fifteen years after the finalisation of the UNGPs, agentic AI presents a remarkable opportunity to stress-test the framework. By mapping the Protect–Respect–Remedy pillars onto the reasoning, acting and interacting capacities of agentic AI systems—and clarifying how companies may cause, contribute to or be linked to harms—organisations can turn high-level principles into practical, layered safeguards across all stages of product development and deployment.

Shared responsibility frameworks, supported by interoperable technical standards, enable the identification, mitigation and ongoing monitoring of human rights risks to be embedded directly into shared infrastructure, rather than addressed only after harm has occurred. Leveraging the UNGPs in this way shifts agentic AI governance from static, point-in-time assessments towards a dynamic model of systemic accountability—one capable not only of preventing harm but also of enabling agentic AI to responsibly advance the realisation of human rights globally.

Cite this article: Angeline Lee and Shahla Naimi, 'Leveraging Human Rights Frameworks for Agentic AI Governance' (2026) *Business and Human Rights Journal,* 1–6. https://doi.org/10.1017/bhj.2026.10055

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 11:46:52, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/bhj.2026.10055