Gartner

Gartner Insights

Responsible AI at Work: Six Principles for CHROs to Mitigate Risk at Scale

LJ Justice, Frank Buytendijk



Responsible AI at Work: Six Principles for CHROs to Mitigate Risk at Scale

28 April 2026 - ID G00852924 - 8 min read

By: LJ Justice, Frank Buytendijk

Initiatives: CxO Leadership; Designing the Enterprise for AI

AI optimizes the workforce but risks scaling discrimination, privacy breaches, and the erosion of employee trust if left ungoverned. Championing AI ethics to protect the workforce is an imperative for CHROs today. Learn how to lead ethical AI adoption using these six foundational principles.

More on This Topic

This is part of an in-depth collection of research. See the collection:

- What CHROs Need to Know About Mitigating AI Talent Risks

Insights at a Glance

CHROs have a unique opportunity to guide the enterprise AI strategy by establishing a dynamic ethical framework that balances AI-driven efficiency with workforce integrity and employee trust.

- AI is embedded in nearly every major workforce decision. Because it relies on historical data to make predictions, it risks scaling existing human prejudices and systemic inequities if left unchecked.
- Managing AI risks is a uniquely human and behavioral challenge. Traditional IT and legal frameworks can miss hidden workforce impacts, such as unjust career outcomes and the erosion of employee trust.
- Rigid one-size-fits-all AI policies foster a dangerous “checklist mentality” and fail to address contextual ethical dilemmas.
- CHROs who ground their organization’s AI strategy in six foundational ethics principles, structuring responsible AI (RAI) as an ongoing process, will mitigate unintended workforce harms and keep human well-being at the center of AI adoption.

Issue

Traditional HR policies stop exactly where modern AI risks begin. Because legacy rulebooks were built exclusively for human-to-human interactions, the rapid integration of AI into everyday processes and workflows creates a dangerous policy gray area.

When it comes to modern AI, particularly machine learning, programmers prepare the datasets and algorithms, but the AI learns by itself. [1] Because AI learns from data from the past to make predictions, it introduces complex, enterprise-scale risks that directly threaten workforce integrity and employee trust. This presents an opportunity for CHROs to govern this shift effectively, guiding the organization's AI strategy, proactively addressing the vulnerabilities associated with the six foundational principles of AI ethics:

1. **Human-centricity risks:** When AI systems are deployed without strict operational boundaries, humans lose the autonomy to apply nuanced judgment to high-stakes decisions — such as those found in healthcare, finance, and autonomous vehicles, where the potential for harm is significantly greater. If systems are allowed to fully automate these processes rather than just augment them, it eliminates the critical human intervention required to prompt reviews and override bad AI recommendations.

2. **Fairness risks:** Because AI trains on historical data, it naturally absorbs past human prejudices and systematic inequities. Without active intervention, AI does not just mirror these historical biases, but is capable of automating and amplifying them at scale. A single flawed model can apply biased judgments to thousands of individuals, unjustly shaping outcomes in lasting ways.

3. **Explainability and transparency risks:** The complex logic behind how an AI model arrives at a specific recommendation is rarely predictable or transparent. If users or affected individuals cannot understand how a decision was made, they cannot contest unfair outcomes, destroying trust and making accountability nearly impossible.

4. **Security and safety risks:** AI models, including generative AI, are trained on massive amounts of data and can inadvertently memorize and expose sensitive personal information or intellectual property. Furthermore, these systems introduce entirely new attack surfaces, making organizations vulnerable to sophisticated phishing, manipulation, and the leakage of confidential data if privacy controls are not integrated.

5. **Accountability risks:** While AI models display basic forms of agency, they cannot assume moral or legal liability. If an algorithmic decision harms an individual or the business, the increasing complexity of these systems obscures clear human ownership, making remediation, liability, and accountability incredibly challenging.

6. **Sustainability risks:** Unmanaged AI deployment demands significant resources, such as massive energy and water consumption, which actively contributes to a larger carbon footprint. The drive for AI-enabled efficiency also risks creating unmanageable productivity expectations that deplete human resilience and cause widespread burnout.

HR leaders need to ask: "Are we managing AI as just another tool, or are we actively governing it as a new, nonhuman actor that is capable of making probabilistic judgments about our employees?"

Impact

As stewards of workforce integrity through fast-evolving AI transformations, CHROs are uniquely equipped to balance corporate control with employee trust. While HR leaders do not code the underlying algorithms, their specific responsibility is to establish the governance mechanisms that mitigate ethical risks and dictate how employees engage with AI responsibly. Because AI is now embedded in nearly every major workforce decision, advancing equitable and defensible outcomes through these tools is an extension of HR's core responsibility.

This is an urgent mandate, as currently, employees cite concerns about AI accuracy and data privacy as the leading barriers preventing them from utilizing enterprise AI tools. [2] Organizations cannot afford to ignore these vulnerabilities, especially as AI is already creating ethical issues in the workplace. For example:

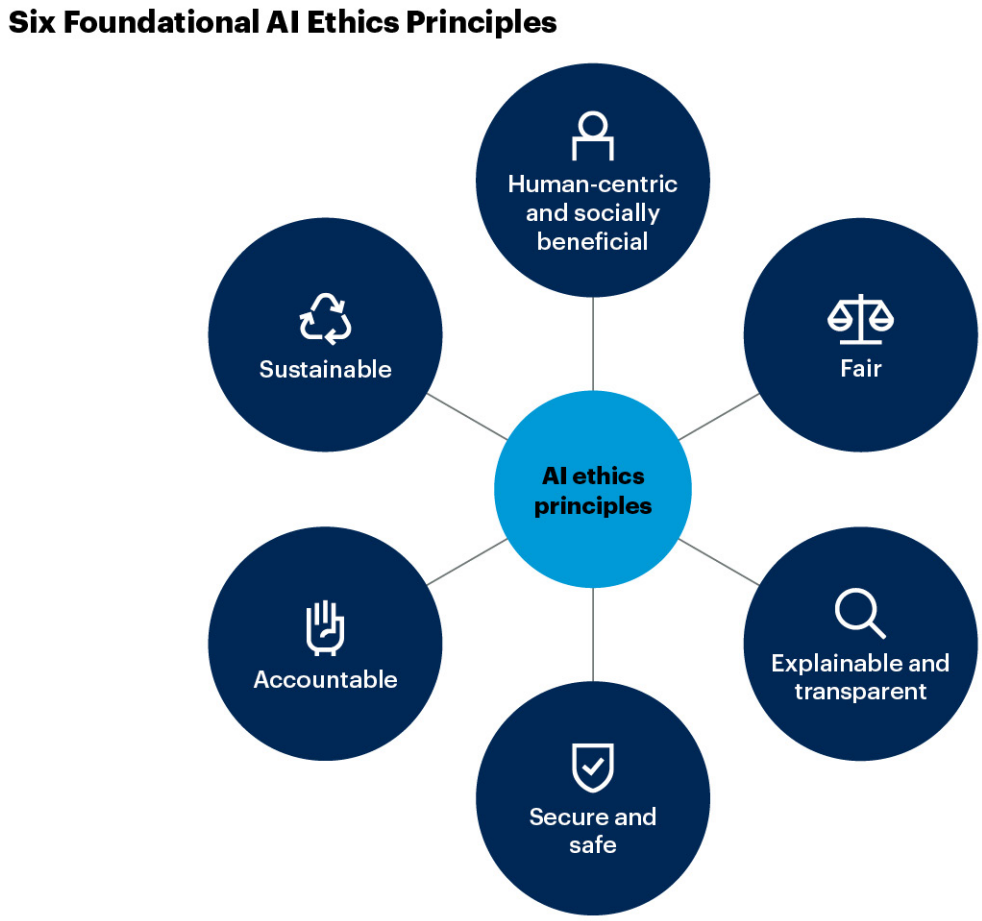
- During tests of AI-enabled HR tools, an AI model told a recruiter that a candidate might struggle to communicate with a "mostly white" team, [3,4]
- Another model dismissed a diverse group of applicants as "diversity tokens," [3,4]
- And in another instance, a major tech company had to scrap its machine learning recruiting tool after discovering it had taught itself to systematically downgrade women's resumes. [5]

These instances demonstrate that when an algorithm inevitably generates a discriminatory or offensive outcome, HR will be held accountable. To get ahead of these risks, CHROs are uniquely positioned to spearhead this effort, guiding their CxO peers in the establishment of dynamic ethical guardrails.

More Detail

The 6 Foundational AI Ethics Principles

Figure 1: Six Foundational AI Ethics Principles



Source: Gartner
822772_C

Gartner

While HR cannot completely immunize the workforce against unintended harm, CHROs can establish the mechanisms and environments needed to actively mitigate these risks. CHROs can lead the charge by collaborating with other CxOs, or establishing a digital ethics advisory board, to anchor the organization's AI strategy in these six foundational principles (see Figure 1). Think of these ethics not as high-level philosophy, but as the operational guardrails necessary to safely govern human-AI work design across the enterprise:

1. **Being human-centric and socially beneficial.** AI solutions should serve human goals, not the other way around. Establish strict boundaries between when AI is allowed to automate a process entirely (for low-risk administrative tasks) and when it is only allowed to augment human judgment (for high-risk, career-impacting decisions). Even when scaling AI, integrate a "human-in-the-loop" approach to enable humans to prompt reviews and override AI recommendations when necessary.

Ask your IT partners or digital ethics advisory board: "Where are we comfortable fully automating a workflow, and where must we mandate human augmentation?"

2. **Being fair.** Because AI adaptively learns from historical data, it easily absorbs and scales deeply embedded social biases. To actively intervene, define the ultimate outcome of fairness: a commitment to impartial treatment and fair access to career opportunities. Rather than making impossible guarantees, fairness requires active human intervention — such as regular bias audits and algorithmic testing — to proactively mitigate the risk of any group being systematically disadvantaged.

Ask your IT partners or digital ethics advisory board: "Are our algorithms inadvertently perpetuating historical inequities or are we actively intervening to correct them?"

3. **Being explainable and transparent.** Employees have a fundamental right to know when and how they are interacting with an AI system, especially if it impacts their careers. Under frameworks like GDPR Article 22, the ability to obtain explanations and contest automated decisions is a legal right. [6] The "black-box" nature of complex models can easily destroy trust. In collaboration with CxOs, mandate that automated decisions across the enterprise remain explainable, testable, and contestable by a human. This requires avoiding opaque manipulation and establishing clear, accessible mechanisms for employees to challenge AI-generated outputs or outcomes.

Ask your IT partners or digital ethics board: "Can our teams and managers actually explain in plain language how this AI model generated its recommendation?"

4. **Being secure and safe.** HR often handles the enterprise's most sensitive personal data. Champion "**privacy by design**," demanding that employee data is masked, encrypted, or replaced with **synthetic data** before being fed into AI training models. Leaders must continuously monitor systems for unintended consequences, such as models drifting or leaking confidential data, all while ensuring that this monitoring does not cross the line into invasive workforce surveillance.

Ask your IT partners or digital ethics advisory board: "Are we protecting employee privacy by design, or treating it as a postimplementation afterthought?"

5. **Being accountable.** AI systems display basic forms of agency, but they cannot assume moral or legal liability. While operational responsibility can be distributed across teams or external vendors, moral and legal accountability cannot be shared with a machine, it will always rest firmly with human leaders. With new regulatory proposals holding deployers liable as developers for system modifications, help the organization establish clear escalation pathways for AI harm, equipping managers, HRBPs, and employees to recognize, route, and remediate flagged incidents.

Ask your IT partners or digital ethics advisory board: "If our AI system makes a discriminatory recommendation today, who exactly is on the hook for fixing it?"

6. **Being sustainable.** AI requires significant energy and water to operate, creating a dual challenge of environmental and social sustainability. Lead the charge in addressing both of these footprints, advocating for proportional technology use and ensuring that AI-driven productivity gains do not translate into unmanageable expectations that deplete workforce resilience and cause widespread employee burnout.

Ask your IT partners or digital ethics advisory board: "Does the actual business value of this AI tool justify its environmental footprint and human toll?"

Evidence

[1] M. Coeckelbergh, "AI Ethics," MIT Press, 2020.

[2] 2026 Gartner Q1 Global Labor Market Survey.

[3] In the “Wild west” of AI chatbots, subtle biases related to race and caste often go unchecked, University of Washington.

[4] “They are uncultured”: Unveiling Covert Harms and Social Threats in LLM-Generated Conversations, ACL Anthology.

[5] Insight — Amazon scraps secret AI recruiting tool that showed bias against women, Reuters.

[6] General Data Protection Regulation (GDPR), GDPR.eu.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

The Cost-Effective Way to Responsible AI: Upskill Your Inclusion Leaders

What to Do When AI Joins the Workforce: Scenarios and HR Policy Starters for CHROs

Ethics, Governance and Compliance Are AI’s Next Frontier

© 2026 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates. This publication may not be reproduced or distributed in any form without Gartner's prior written permission. It consists of the opinions of Gartner's Business and Technology Insights Organization, which should not be construed as statements of fact. While the information contained in this publication has been obtained from sources believed to be reliable, Gartner disclaims all warranties as to the accuracy, completeness or adequacy of such information. Although Gartner insights may address legal and financial issues, Gartner does not provide legal or investment advice and its insights should not be construed or used as such. Your access and use of this publication are governed by Gartner's Usage Policy. Gartner prides itself on its reputation for independence and objectivity. Its insights is produced independently by its Business and Technology Insights Organization without input or influence from any third party. For further information, see " Guiding Principles on Independence and Objectivity." Gartner insights may not be used as input into or for the training or development of generative artificial intelligence, machine learning, algorithms, software, or related technologies.

Actionable, objective insights

Position your organization for success. Explore these additional complimentary resources and tools.



Report

The CHRO Report

Quarterly insights for HR leaders on challenges, risks and actions.

Download Now



Webinar

The Top CHRO Trends and Priorities Shaping 2027

Expert insights and recommendations to shape HR's strategy for 2027.

Watch Now



Roadmap

AI Roadmap for HR

Discover key steps to build an AI roadmap that delivers value in HR.

Download Now



Guide

CHRO Guide for HR Transformation

View practical strategies to reinvent HR and deliver value in the AI era.

Download Now

Already a client? Get access to even more resources in your client portal.
Log In ↗

Connect with us

Get actionable, objective business and technology insights that drive smarter decisions and stronger performance on your mission-critical priorities. Contact us to become a client:

U.S.: 1 855 322 5484
International: +44 (0) 3300 296 946

Become a Client

Learn more about Gartner for CHROs
gartner.com/en/human-resources

Stay connected to the latest insights

Attend a Gartner conference
View Conference

© 2026 Gartner, Inc. and/or its affiliates. All rights reserved. CM_GBS_5497850

Gartner®