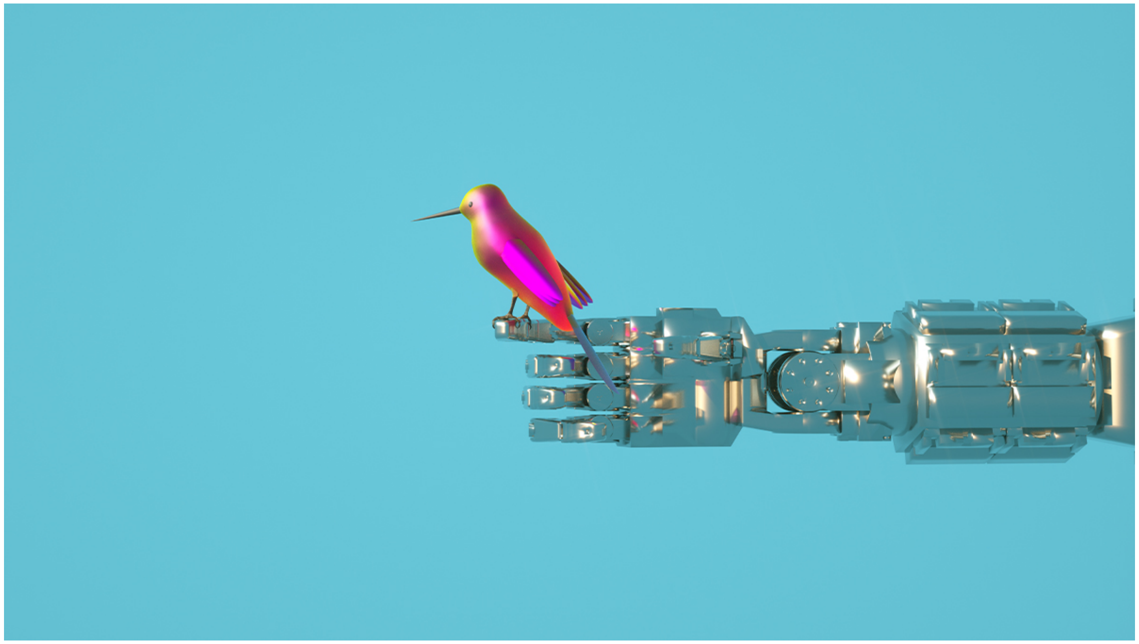
Harvard
Business
Review

Technology And Analytics

13 Principles for Using AI Responsibly

by Brian Spisak, Louis B. Rosenberg, and Max Beilby

June 30, 2023



Boris SV/Getty Images

Summary. The competitive nature of AI development poses a dilemma for organizations, as prioritizing speed may lead to neglecting ethical guidelines, bias detection, and safety measures. Known and emerging concerns associated with AI in the workplace include the spread of misinformation, copyright and intellectual property concerns, cybersecurity, data privacy, as well as navigating rapid and ambiguous regulations. To mitigate these risks, we propose thirteen principles for responsible AI at work. **close**

Love it or loath it, the rapid expansion of AI will not slow down anytime soon. But AI blunders can quickly damage a brand's reputation — just ask Microsoft's first chatbot, Tay. In the tech race, all leaders fear being left behind if they slow down while others don't. It's a high-stakes situation where cooperation seems risky, and defection tempting. This "prisoner's dilemma" (as it's called in game theory) poses risks to responsible AI practices. Leaders, prioritizing speed to market, are driving the current AI arms race in which major corporate players are rushing products and potentially short-changing critical considerations like ethical guidelines, bias detection, and safety measures. For instance, major tech corporations are laying off their AI ethics teams precisely at a time when responsible actions are needed most.

It's also important to recognize that the AI arms race extends beyond the developers of large language models (LLMs) such as OpenAI, Google, and Meta. It encompasses many companies utilizing LLMs to support their own custom applications. In the world of professional services, for example, PwC announced it is deploying AI chatbots for 4,000 of their lawyers, distributed across 100 countries. These AI-powered assistants will “help lawyers with contract analysis, regulatory compliance work, due diligence, and other legal advisory and consulting services.” PwC’s management is also considering expanding these AI chatbots into their tax practice. In total, the consulting giant plans to pour $1 billion into “generative AI” — a powerful new tool capable of delivering game-changing boosts to performance.

In a similar vein, KPMG launched its own AI-powered assistant, dubbed KymChat, which will help employees rapidly find internal experts across the entire organization, wrap them around incoming opportunities, and automatically generate proposals based on the match between project requirements and available talent. Their AI assistant “will better enable cross-team collaboration and help those new to the firm with a more seamless and efficient people-navigation experience.”

Slack is also incorporating generative AI into the development of Slack GPT, an AI assistant designed to help employees work smarter not harder. The platform incorporates a range of AI capabilities, such as conversation summaries and writing assistance, to enhance user productivity.

These examples are just the tip of the iceberg. Soon hundreds of millions of Microsoft 365 users will have access to Business Chat, an agent that joins the user in their work, striving to make sense of their Microsoft 365 data. Employees can prompt the assistant to do everything from developing status report summaries based on meeting transcripts and email communication to identifying flaws in strategy and coming up with solutions.

This rapid deployment of AI agents is why Arvind Krishna, CEO of IBM, recently wrote that, “[p]eople working together with trusted A.I. will have a transformative effect on our economy and society … It’s time we embrace that partnership — and prepare our workforces for everything A.I. has to offer.” Simply put, organizations are experiencing exponential growth in the installation of AI-powered tools and firms that don’t adapt risk getting left behind.

AI Risks at Work

Unfortunately, remaining competitive also introduces significant risk for both employees and employers. For example, a 2022 UNESCO publication on “the effects of AI on the working lives of women” reports that AI in the recruitment process, for example, is excluding women from upward moves. One study the report cites that included 21 experiments consisting of over 60,000 targeted job advertisements found that “setting the user’s gender to ‘Female’ resulted in fewer instances of ads related to high-paying jobs than for users selecting ‘Male’ as their gender.” And even though this AI bias in recruitment and hiring is well-known, it’s not going away anytime soon. As the UNESCO report goes on to say, “A 2021 study showed evidence of job advertisements skewed by gender on Facebook even when the advertisers wanted a gender-balanced audience." It's often a matter of biased data which will continue to infect AI tools and threaten key workforce factors such as diversity, equity, and inclusion.

Discriminatory employment practices may be only one of a cocktail of legal risks that generative AI exposes organizations to. For example, OpenAI is facing its first defamation lawsuit as a result of allegations that ChatGPT produced harmful misinformation. Specifically, the system produced a summary of a real court case which included fabricated accusations of embezzlement against a radio host in Georgia. This highlights the negative impact on organizations for creating and sharing AI generated information. It underscores concerns about LLMs fabricating false and libelous content, resulting in reputational damage, loss of credibility, diminished customer trust, and serious legal repercussions.

In addition to concerns related to libel, there are risks associated with copyright and intellectual property infringements. Several high-profile legal cases have emerged where the developers of generative AI tools have been sued for the alleged improper use of licensed content. The presence of copyright and intellectual property infringements, coupled with the legal implications of such violations, poses significant risks for organizations utilizing generative AI products. Organizations can improperly use licensed content through generative AI by unknowingly engaging in activities such as plagiarism, unauthorized adaptations, commercial use without licensing, and misusing Creative Commons or open-source content, exposing themselves to potential legal consequences.

The large-scale deployment of AI also magnifies the risks of cyberattacks. The fear amongst cybersecurity experts is that generative AI could be used to identify and exploit vulnerabilities within business information systems, given the ability of LLMs to automate coding and bug detection, which could be used by malicious actors to break through security barriers. There's also the fear of employees accidentally sharing sensitive data with third-party AI providers. A notable instance involves Samsung staff unintentionally leaking trade secrets through ChatGPT while using the LLM to review source code. Due to their failure to opt out of data sharing, confidential information was inadvertently provided to OpenAI. And even though Samsung and others are taking steps to restrict the use of third-party AI tools on company-owned devices, there's still the concern that employees can leak information through the use of such systems on personal devices.

On top of these risks, businesses will soon have to navigate nascent, varied, and somewhat murky regulations. Anyone hiring in New York City, for instance, will have to ensure their AI-powered recruitment and hiring tech doesn't violate the City's "automated employment decision tool" law. To comply with the new law, employers will need to take various steps such as conducting third-party bias audits of their hiring tools and publicly disclosing the findings. AI regulation is also scaling up nationally with the Biden-Harris administration's "Blueprint for an AI Bill of Rights" and internationally with the EU's AI Act, which will mark a new era of regulation for employers.

This growing nebulous of evolving regulations and pitfalls is why thought leaders such as Gartner are strongly suggesting that businesses "proceed but don't over pivot" and that they "create a task force reporting to the CIO and CEO" to plan a roadmap for a safe AI transformation that mitigates various legal, reputational, and workforce risks. Leaders dealing with this AI dilemma have important decision to make. On the one hand, there is a pressing competitive pressure to fully embrace AI. However, on the other hand, a growing concern is arising as the implementation of irresponsible AI can result in severe penalties, substantial damage to reputation, and significant operational setbacks. The concern is that in their quest to stay ahead, leaders may unknowingly introduce potential time bombs into their organization, which are poised to cause major problems once AI solutions are deployed and regulations take effect.

For example, the National Eating Disorder Association (NEDA) recently announced it was letting go of its hotline staff and replacing them with their new chatbot, Tessa. However, just days before making the transition, NEDA discovered that their system was promoting harmful advice such as encouraging people with eating disorders to restrict their calories and to lose one to two pounds per week. The World Bank spent $1 billion to develop and deploy an algorithmic system, called Takaful, to distribute financial assistance that Human Rights Watch now says ironically creates inequity. And two lawyers from New York are facing possible disciplinary action after using ChatGPT to draft a court filing that was found to have several references to previous cases that did not exist. These instances highlight the need for well-trained and well-supported employees at the center of this digital transformation. While AI can serve as a valuable assistant, it should not assume the leading position.

Principles for Responsible AI at Work

To help decision-makers avoid negative outcomes while also remaining competitive in the age of AI, we've devised several principles for a sustainable AI-powered workforce. The principles are a blend of ethical frameworks from institutions like the National Science Foundation as well as legal requirements related to employee monitoring and data privacy such as the Electronic Communications Privacy Act and the California Privacy Rights Act. The steps for ensuring responsible AI at work include:

- **Informed Consent.** Obtain voluntary and informed agreement from employees to participate in any AI-powered intervention *after* the employees are provided with all the relevant information about the initiative. This includes the program's purpose, procedures, and potential risks and benefits.
- **Aligned Interests.** The goals, risks, and benefits for both the employer and employee are clearly articulated and aligned.

- **Opt In & Easy Exits.** Employees must opt into AI-powered programs without feeling forced or coerced, and they can easily withdraw from the program at any time without any negative consequences and without explanation.
- **Conversational Transparency.** When AI-based conversational agents are used, the agent should formally reveal any persuasive objectives the system aims to achieve through the dialogue with the employee.
- **Debiased and Explainable AI.** Explicitly outline the steps taken to remove, minimize, and mitigate bias in AI-powered employee interventions—especially for disadvantaged and vulnerable groups—and provide transparent explanations into how AI systems arrive at their decisions and actions.
- **AI Training and Development.** Provide continuous employee training and development to ensure the safe and responsible use of AI-powered tools.
- **Health and Well-Being.** Identify types of AI-induced stress, discomfort, or harm and articulate steps to minimize risks (e.g., how will the employer minimize stress caused by constant AI-powered monitoring of employee behavior).
- **Data Collection.** Identify what data will be collected, if data collection involves any invasive or intrusive procedures (e.g., the use of webcams in work-from-home situations), and what steps will be taken to minimize risk.

- **Data.** Disclose any intention to share personal data, with whom, and why.
- **Privacy and Security.** Articulate protocols for maintaining privacy, storing employee data securely, and what steps will be taken in the event of a privacy breach.
- **Third Party Disclosure.** Disclose all third parties used to provide and maintain AI assets, what the third party's role is, and how the third party will ensure employee privacy.
- **Communication.** Inform employees about changes in data collection, data management, or data sharing as well as any changes in AI assets or third-party relationships.
- **Laws and Regulations.** Express ongoing commitment to comply with all laws and regulations related to employee data and the use of AI.

We encourage leaders to urgently adopt and develop this checklist in their organizations. By applying such principles, leaders can ensure rapid *and* responsible AI deployment.

Brian R. Spisak *is a Research Associate at Harvard's National Preparedness Leadership Initiative and an independent consultant. He is the author of* Computational Leadership: Connecting Behavioral Science and Technology

to Optimize Decision-Making and Increase Profits (Wiley, 2023).

Louis B. Rosenberg *is the CEO and Chief Scientist of Unanimous AI, the Chief Scientist of the Responsible Metaverse Alliance, and a former professor at California State University (Cal Poly). He was previously Founder and CEO of Immersion Corporation (IMMR Nasdaq) and Outland Research.*

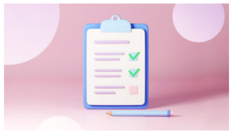
MB

Max Beilby *is a business psychologist, working in financial services. Max authors the blog Darwinian Business and is a Senior Vice President in Citi's Culture & Behavioral Risk team. Max is also a member of Ethical Systems' Advisory Board.*

Recommended For You

28 Questions to Ask Your Boss in Your One-on-Ones

10 Common Job Interview Questions and How to Answer Them



38 Smart Questions to Ask in a Job Interview

PODCAST

How to Manage: Finding Yourself Again