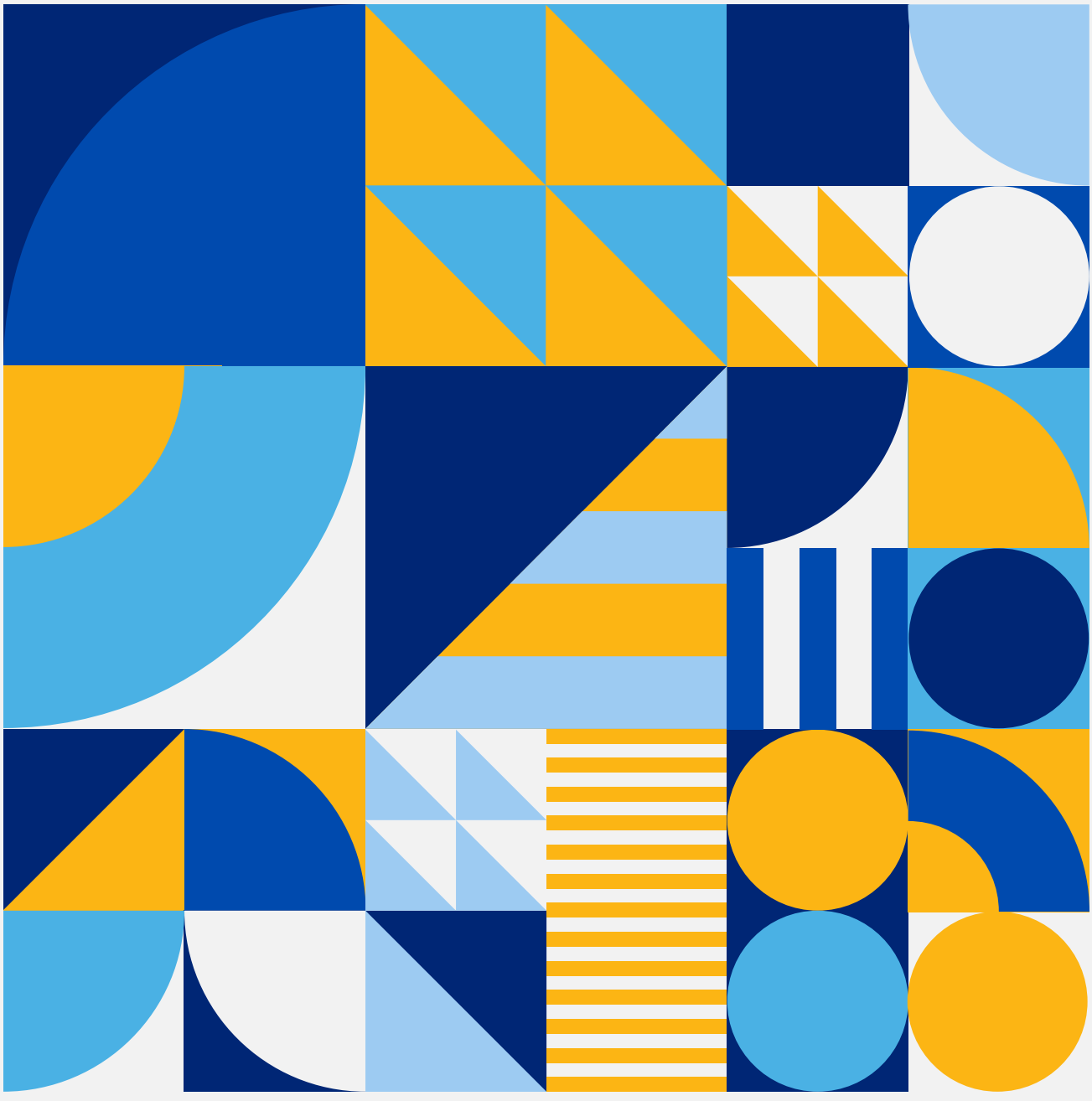


Responsible Use of Generative AI

A Playbook for Product Managers & Business Leaders

BAIR Responsible AI Initiative

Berkeley Haas

Responsible use of generative AI means proactively addressing potential risks and harms to embed trust and foster accountability.

What is This Playbook?

This playbook includes 10 plays for product managers and business leaders to use generative AI (genAI) responsibly— including in day-to-day work and in new products or services.

Who is This Playbook For?

This playbook is for product managers who are using genAI in their day-to-day work and in new products. It is also for organizational decision makers who are grappling with adoption of genAI tools in their workplace and products or services.

Why Use This Playbook?

In order to unlock the full potential of generative AI (genAI), it is important to address its risks and ensure genAI is used 'responsibly'.

How to Use This Playbook?

The playbook starts with an overview of the current state of genAI adoption in workplaces, challenges in using genAI responsibly, what responsibility means or looks like, and the business case for responsible use of genAI. It then explores each play which includes: who is involved, how to implement, case studies, as well as tools and resources.

How and by whom was this playbook developed?

This Playbook was authored by Genevieve Smith (University of California (UC) Berkeley), Natalia Luka (UC Berkeley), Jessica Newman (UC Berkeley), Merrick Osborne (UC Berkeley), Brandie Nonnecke (UC Berkeley), Brian Lattimore (Stanford University), and Brent Mittelstadt (University of Oxford). The playbook builds off a research project led by the Responsible AI Initiative of the Berkeley AI Research Lab with Berkeley Haas and conducted by a team that resulted in an academic paper, "Responsible Generative AI Use by Product Managers: Recoupling Ethical Principles and Practices". The playbook was prototyped with product managers. The project received funding support from Google. (See Appendix for full acknowledgements)

Roadmap

Executive Summary[1]

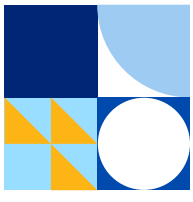
This playbook focuses on the responsible use of Generative AI (genAI) for product managers. Using genAI responsibly entails proactively addressing potential risks and harms thereby embedding trust and fostering accountability.

Key Understandings:

1. GenAI is being adopted rapidly across industries for day-to-day work and products.

- **Diverse applications:** GenAI is being used for tasks like automating work, generating content, transcribing voice, and powering new products and features.
- **Model options:** Organizations leverage genAI through off-the-shelf tools, enterprise solutions, or by leveraging more open models to customize and tailor to specific needs and products.
- **Benefits:** Adoption can lead to productivity and efficiency gains, with value creation varying by business function. Organizations benefiting most are paying attention to genAI risks, while those lagging in addressing risks are inhibited from capitalizing on benefits.

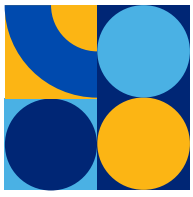
2. Those who are and will continue to win from using genAI, are practicing responsibility. There is a clear business case.



- **Builds trust & brand reputation:** Responsible AI practices foster positive brand image and customer loyalty.
- **Maintains regulatory compliance:** Proactive mitigation minimizes risks associated with evolving AI regulations.
- **Mitigates risk & drives sustainable growth:** Addressing ethical concerns supports long-term value creation and avoids reputational damage or legal penalties.

[1] *This executive summary was developed with the assistance of NotebookLM.*

3. There are key risks in using genAI—and particularly five—that Product Managers need to pay attention to: data privacy, transparency, inaccuracy, bias, safety and security.



- **Data Privacy:** GenAI models may retain user data, raising concerns about long-term privacy and potential copyright infringement. Unintened data exposure is possible, with models potentially revealing personal information or copyrighted material from their training data.
- **Transparency:** The "black box" nature of genAI models makes it difficult to understand how decisions are made and why certain outputs were produced. Meanwhile, companies developing genAI systems are often not practicing transparency, withholding details about training data, model architectures, and decision-making processes.
- **Hallucinations & Inaccuracy:** GenAI tools are known to confidently assert false information or "hallucinate," impacting their usefulness and trustworthiness.
- **Bias:** Gen AI models can exhibit biases based on training data. This can include performing worse for certain populations or groups, and reinforcing harmful stereotypes or discrimination.
- **Safety & Security:** Vulnerabilities, like prompt injection attacks, can cause data leaks or provision of dangerous information.
- Additional concerns exist about the **future of work**, **environmental impacts**, and **copyright infringement**.

4. There are several challenges to using genAI responsibly.

- Lack of **organizational** policies and approaches coupled with misaligned incentives and lack of individual education.
- General immaturity in the **industry** as it relates to responsibility.
- The replication and reinforcement of inequitable patterns that exist in **society**.

Using gen AI responsibly takes action both at the Organizational Leadership level and the individual Product Manager level.

There are 5 plays for **Organizational Leaders** and 5 plays for **Product Managers**:

Organizational Leader Plays: *These plays focus on integrating responsibility across the organization by aligning leadership, governance, policies, and culture to drive accountability and trust.*

1. Ensure leadership recognizes the value of responsible genAI use, develop responsible AI principles, and communicate the organization's commitment to responsibility to all employees.
2. Implement policies and accompanying standards to ensure responsible use of generative AI.
3. Build a comprehensive responsible AI governance framework that defines key roles, establishes organizational structures, and fosters a culture of shared accountability.
4. Update incentives to align performance, product development, and metrics with responsibility.
5. Implement tailored training to address gaps and support responsible use of genAI.

Product Manager Plays: *These plays outline practical actions for regular responsibility practices to ensure trustworthy use and product development.*

1. Conduct "gut checks" to evaluate responsibility risks in work use cases and product development.
2. Choose a model for genAI products by assessing needs and potential risks. Ensure transparency by documenting the model, fine-tuning data, and key considerations.
3. Conduct risk assessments and audits for genAI products, involving cross-functional teams, expert oversight, and tools aligned with organizational principles and core risks.
4. Implement red-teaming and adversarial testing to uncover vulnerabilities, while capturing and responding to user feedback over time.
5. Track your responsibility micro-moments—simple, impactful actions that demonstrate responsible decision-making—and showcase them in performance reviews.

Overall, this playbook provides a comprehensive guide for navigating the landscape of responsible genAI use. It equips you with the knowledge, strategies, and tools necessary to harness the power of genAI while mitigating its potential risks and ensuring ethical and sustainable implementation.

Choose Your Playbook Path:

Are you an Organizational Leader?

→ **YES**

Are you newer to genAI?

✔ Yes! I am not very familiar with what genAI is, how it is being adopted by organizations, and the risks that exist.

→ **Start at the beginning**

Are you familiar with genAI and the responsibility risks?

✔ I'm familiar with genAI *(including what genAI is and the ways it is being adopted by organizations)*, but I am **NOT** familiar with the risks the technology poses.

→ **Learn about the risks before going to the plays**

✔ I'm familiar with genAI *(including what it is and how it's being adopted by organizations)*. I'm also familiar with the risks the technology poses.

→ **Go to the Organizational Leader plays**

or

Are you a Product Manager?

→ **YES**

Are you newer to genAI?

✔ Yes, I'm a newbie to this space! I am not very familiar with what genAI is, how it is being adopted by organizations, and the risks that exist.

→ **Start at the beginning**

Are you familiar with genAI and the responsibility risks?

✔ I'm familiar with genAI *(including what genAI is and the ways it is being adopted by organizations)*, but I am **NOT** familiar with the risks the technology poses.

→ **Learn about the risks before going to the plays**

✔ I'm familiar with genAI *(including what it is and how it's being adopted by organizations)*. I'm also familiar with the risks the technology poses.

→ **Go to the Product Manager plays**

You've read the playbook. Now what?

1. **Make a list of plays relevant for you and your organization.** Start putting them into action, following the guidance and leveraging the resources provided.
2. **Build leadership support** for broader responsibility efforts. You know your company and context best, but here are some ideas to gather internal support:
   - ☐ Highlight the business case for responsible use of genAI.
   - ☐ Use examples in your industry and application(s) of AI where responsible use unlocked new value, or where irresponsible use led to costly avoidable mistakes.
   - ☐ Connect responsible genAI use to the company's values and AI principles (if they exist)
   - ☐ Link / connect the importance of responsible use of genAI to achieving specific OKRs (Objectives and Key Results)
3. **Reflect on and revisit your own progress on the plays.** Share them with others.

I. Introduction

Sarah[2], a product manager juggling tight deadlines and endless feature requests at her fintech company, just began experimenting with genAI to ease her workload. She used it to draft user stories, summarize market trends, and brainstorm product names—but when the market summaries misrepresented key trends and the product names inadvertently echoed cultural stereotypes it sparked a nagging worry about accuracy and bias. As her company explored integrating AI into customer-facing products, Sarah wrestled with questions around trustworthiness, wondering how to wield this powerful tool without compromising her values or the company's reputation.

Following the launch of ChatGPT in November 2022, usage of generative AI has exploded across organizations globally. GenAI is an incredibly helpful tool for a variety of daily work use cases—from automating work tasks like coding, writing support, data analysis, and more. Meanwhile, entrepreneurs and product managers like Sarah recognize the immense innovation and economic opportunities genAI opens when integrated into new products and features. Now, in 2025, the hype continues—not just being urged on by innovators, companies, and investors, but countries and governments eager to lead and capitalize on the technology.

There are numerous benefits of genAI. The technology can help people make decisions more efficiently and cost-effectively, while promoting higher productivity and business growth. While the use of AI in predictions and decision making can reduce human subjectivity and open new opportunities: it also opens up potential for bias, inaccuracies, data privacy violations, and more.

In order to unlock the full potential of genAI, organizations are increasingly recognizing the importance of addressing these risks and ensuring genAI is used 'responsibly'. But what does 'responsible' use of genAI in day-to-day work and new products mean and look like?

This playbook, built from academic research, is for product managers who are using GenAI in their work and in new products. It is also for organizational decision makers who are grappling with adoption of genAI tools in their workplace and products or services.

The playbook outlines key plays—for business leaders and for individual product managers—that outline how to responsibly use genAI in day-to-day work and in new products. Before delving into the plays, it provides an overview of the current state of genAI adoption in workplaces, challenges in using genAI responsibly, what responsibility means or looks like, and the business case for responsible use of genAI.

The goal of this playbook is to help you and your business capitalize on genAI while ensuring responsibility and advancing trust across employees, customers, and society more broadly.

[2]*Sarah is fictional, but based on real people and ways that product managers are using genAI (building from our interviews with product managers in a range of industries)*

II. Background

a. What is generative AI?

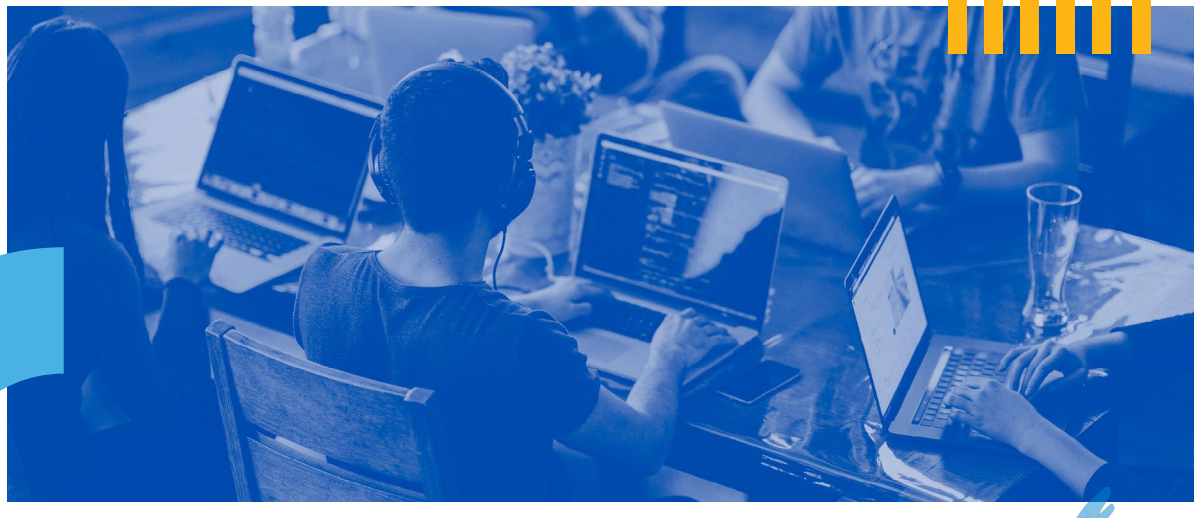
Many of the AI systems used by organizations use machine learning (ML), in which a series of algorithms takes and learns from massive amounts of data to find patterns and make predictions. There are two types of ML models: discriminative and generative. Discriminative models classify or predict (e.g., who should be prioritized for a vaccine shot, who should get a job interview from a hiring pool, amount of credit to offer an individual). **Generative models generate new data—including text, code, images, video, and more.**[3]

GenAI tools are often built from foundation models, which are models trained on massive datasets and based on complex neural networks. Foundation models use learned patterns and relationships to predict the next item in a sequence. They are different from traditional ML models, due to their size and general-purpose nature (as opposed to ML models that may perform specific tasks like classifying images). Foundation models use self-supervised learning, meaning the models do not learn from labeled training datasets, but create their own labels from the input data. They can continue learning from data inputs or prompts. They are costly to develop and maintain, but their size and flexibility enable a wide range of applications.

Popular foundation models include OpenAI's GPT-4, Google's Gemini, Anthropic's Claude, and Meta's Llama 3. Many are large language models (LLMs) trained to generate text, while multimodal models process text, video, audio, and images.

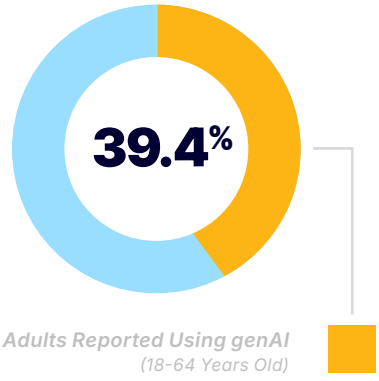
These models exist on an openness spectrum ranging from open source to fully closed. Model developers decide whether to make each component of the training, evaluation and deployment pipeline private (closed) or public (open), with varying levels of restrictions for the latter. While there are positives and negatives across the openness spectrum, open models are considered more flexible and customizable as they allow developers to have access to more training approaches, models and datasets that enable users to tailor models to their use case and application.[4] They also provide more transparency and give greater control of the data pipeline. Closed models accessed via an API make product developers reliant on an external provider for key aspects of the product or system that can limit control and maintainability, but can offer easier integration.

Foundation models act as engines for specialized downstream applications, accelerating and lowering the cost of new ML developments. For instance, GPT-4 powers ChatGPT.[5][6] These models and generative AI tools are transforming industries by generating content, automating tasks, transcribing voice, and more.

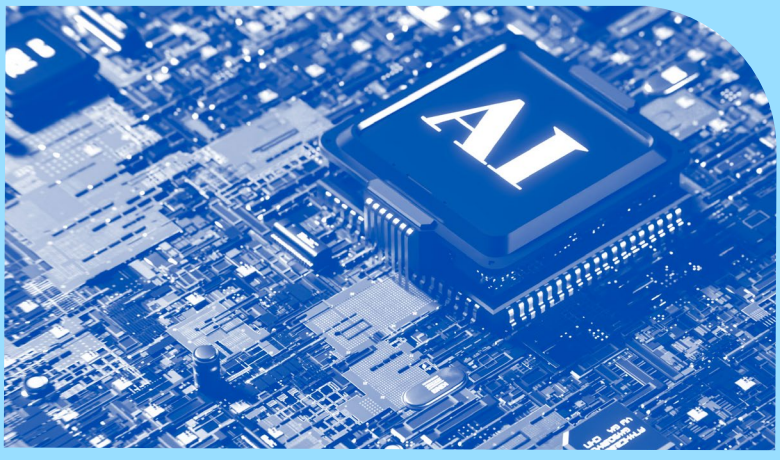


b. How much, by whom, and for what purposes is genAI being adopted?

AI is being adopted rapidly. Recent research finds that 39.4% of US adults (18-64 years old) reported using genAI, with 24% of workers using it at least once a week and 11% daily—across a range of occupations and tasks.[7] Adoption has increased rapidly: A 2024 McKinsey study similarly finds a near doubling of genAI use across all regions in the past year. ChatGPT, a leading genAI tool, boasts 200 million weekly active users as of August 2024, double the number from 2023.[8] The pace of adoption for genAI is quicker than the adoption pace for both PCs and the Internet.[9]



US Adults Reported Using genAI (18-64 Years Old)

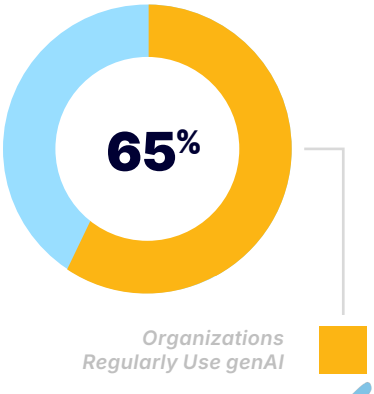


Adoption by role & industry:

GenAI adoption is highest among senior leaders and workers in professional services, energy, and materials sectors. Nearly half of the workers adopting genAI are in computer, math, and management roles.[10] The most common job tasks it is used for are: writing; admin support; interpreting, translating and summarizing text or data; coding; documentation or detailed instructions; and idea generation. The most common products are OpenAI's ChatGPT, followed by Google's Gemini, then embedded products.[11]

Organizational adoption:

Organizations are also integrating genAI technologies at unprecedented rates into their operations and new products. A 2024 McKinsey Survey found that 65% of organizations regularly use genAI, nearly double from just ten months prior.[12] Adoption is global: over two-thirds of respondents in nearly every region say their organizations are using some type of AI. As of August 2024, OpenAI states that 92% of Fortune 500 companies now use its tools.[13]



Organizations Regularly Use genAI

Business applications:

Within organizations, the most common business function is marketing and sales (34% reporting regular genAI use) and product / service development (23%).[14] Within product / service development, the most common genAI use cases are design development, literature and research review, and early testing. Many organizations are also integrating genAI into products and features. For example, Canva integrated OpenAI technology into its feature, Magic Write[15], and used Stable Diffusion (an image-generation model) to create its Magic Media[16] tool. GPT-4 was used to create Spotify's AI DJ.[17] Foundation models can also be used to develop powerful internal organizational tools in areas like HR or customer service, driving efficiencies in processes such as recruitment, onboarding, and client support.

c. Off-the-Shelf, Customizing, or Developing Own Models?

Organizations are leveraging generative AI through off-the-shelf tools, enterprise solutions, procurement contracts, and custom models. Off-the-shelf tools like OpenAI's ChatGPT and Google's Gemini are popular for rapid deployment with minimal setup, especially in industries like business, legal, and professional services. In contrast, sectors like energy and materials are more likely to develop custom or extensively fine-tuned models to meet their unique requirements.[18]

Enterprise solutions—such as OpenAI's ChatGPT Enterprise, Azure-powered offerings, or Anthropic's Claude for Business—provide additional capabilities tailored for organizations. These licenses allow companies to fine-tune models with proprietary data for specific tasks like customer service, coding, or document summarization, while protecting data privacy through encryption and isolated data pipelines. Enterprise models also support customization without direct fine-tuning, using features like embeddings or dynamic prompt engineering. For instance, PwC developed ChatPwC, an internal tool built on Microsoft's Azure OpenAI Service, customized with proprietary PwC data.[19] Lastly, procurement can allow organizations to source genAI solutions that meet their particular needs and can result in longer-term contracts and partnerships.

What about for new products or features?

When developing new products, product managers can choose among pre-trained closed models, enterprise models, and more open models, depending on their requirements for control, flexibility, and scalability:

- **Pre-trained closed / off-the-shelf models** (accessed via APIs), such as Anthropic's Claude, allow for rapid prototyping and easy integration with minimal technical overhead. They can be helpful for early stage development, but have limited flexibility and rely on external providers.
- **Enterprise models** allow for added control, security, and privacy protections. They often support fine-tuning and customization for specific applications, but can be costly and still dependent on external providers.
- **More open models**, such as Meta's Llama, provide control and flexibility. These models enable fine-tuning on proprietary data, custom architecture adjustments, and modification of training pipelines. However, they require significant technical expertise and resources for deployment and scaling.

d. Adoption of genAI

There are immense benefits of genAI. The technology can help people make decisions more efficiently and cost-effectively, while promoting higher productivity and economic growth.[20] Use of AI in predictions and decision making can reduce human subjectivity and open new opportunities.

GenAI can lead to productivity and efficiency gains. Researchers found that giving GPT-4 access to Boston Consulting Group consultants led to significant productivity gains including completing 12.2% more tasks on average, completing them 25.1% more quickly, and produced higher quality results.[21] Importantly, for certain topics and activities the AI system was not sufficiently capable and led to reductions in quality.

Researchers estimate that between .5 and 3.5% of all work hours in the US are currently supported by genAI, which could potentially boost labor productivity .125 to .875 percentage points at current usage levels.[22]

Value creation by genAI varies by business function. In a survey of global executives by McKinsey, 39% of respondents report cost decreases across all functions.[23] Human resources (HR) reported the largest cost decreases (50% of respondents report decreases followed by supply chain and inventory management, service operations, IT and software engineering). The same survey found that 44% of respondents report meaningful revenue increases.[24] Risk, legal, and compliance are more likely to report meaningful revenue increases, followed by IT, marketing and sales.

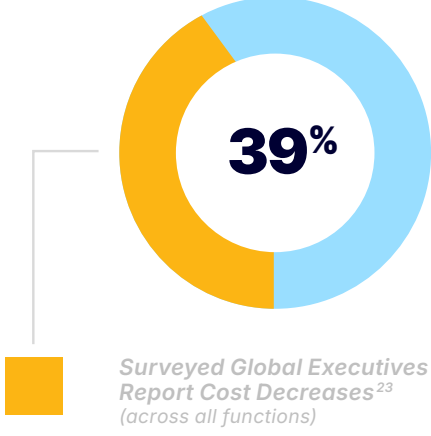
Adoption Not Ubiquitous

Not all organizations have been eager to adopt genAI. There are various challenges facing adoption related to issues and responsibility risks with genAI, governance and regulatory uncertainty, complicated architectural challenges, the sprawl of AI choices, costs for implementation, and security vulnerabilities.

In May 2023, Apple restricted its employees from using genAI tools including ChatGPT and CoPilot due to concerns about potential leaks of confidential data.[25] Other companies that have banned or restricted employee use of ChatGPT and other genAI tools include Spotify, Verizon, Wells Fargo, Samsung, Deutsche Bank, and Amazon, among others.[26] Many organizations worry about employees inadvertently providing sensitive or proprietary information to these models without proper safeguards.

In some cases, companies have modified restrictions and embraced genAI. For example, in June 2024, Apple and OpenAI announced a partnership, which includes integrating ChatGPT into Apple's iOS, iPadOS, and macOS. Apple users will be able to access ChatGPT through Siri and across Apple's Writing Tools.[27]

In other instances, organizations have stopped or limited their use of genAI tools due to the production of inaccurate information or biased and harmful content that puts the company at risk of liability and reputational harm. For example, Air Canada was using an AI chatbot to provide answers to customers about its policies. The chatbot told one passenger that he was eligible for getting a bereavement fare discount after booking a fare, but the company refused to follow through with this, stating that the chatbot had been wrong



Surveyed Global Executives Report Cost Decreases[23]
(across all functions)

about their policy and claiming that the chatbot was a “separate legal entity that is responsible for its own actions.”[28] The passenger filed a claim against the company with the British Columbia Civil Resolution Tribunal, which rejected Air Canada’s argument and held them liable for the chatbot’s faulty advice. High profile stories such as these—and the risks they represent—make companies wary of integrating genAI into public-facing tools.

But genAI high performers are excelling—including by addressing responsibility concerns

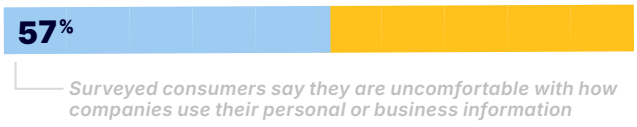
It is still the early days of adoption and only a small number of organizations are attributing a meaningful share of Earnings Before Interest and Taxes (EBIT) to deployment of AI (46 of 876 global executives responding to a McKinsey 2024 survey).[29] These organizations, considered ‘genAI high performers’ by McKinsey, are using genAI in more business functions (i.e. in three functions including most often marketing and sales and product and service development, as well as one other area). **In addition to adoption in several business functions, these genAI high performers are paying more attention to genAI related risks and follow a set of risk-related best practices.**

Many organizations are lagging in addressing risks and developing robust responsible AI (RAI) approaches, which inhibit their ability to capitalize on benefits. A study of RAI “maturity" across 1000 organizations in 20 industries and 19 geographical regions found that *"the majority of organizations are at mid-level organizational RAI maturity, which can be interpreted as an indication of a broad recognition of RAI while highlighting challenges in advancing beyond this stage"*.[30] The study found that **only 9% of organizations have achieved "Optimized" responsible AI organizational maturity, and only 0.8% have reached operational maturity.** The gap between having an RAI practice on paper and implementing it means that organizations can appear more prepared to handle AI responsibly than they actually are. This can lead to a false sense of security from customers, or more trust in an organization/ service than is warranted. In short, fully capitalizing on AI requires filling this gap to implement robust responsibility approaches that account for the various responsibility risks present in genAI.

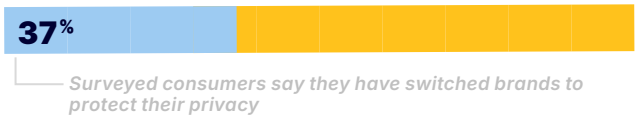
III. What is the business case for responsible use of genAI?

Using genAI responsibly is smart business. Responsible use of genAI can build trust and strong brand reputation, and maintain regulatory compliance and avoid costly changes, while mitigating risks and driving sustainable growth.

Build Trust & Brand Reputation: Responsible AI practices enhance stakeholder trust, fostering a positive brand image and customer loyalty. An IBM survey finds that 57% of consumers say they are uncomfortable with how companies use their personal or business information and 37% have switched brands to protect their privacy.[31] Meanwhile, RAI practices such as enhanced transparency increase trust.[32] Moreover, responsible practices help prevent reputational damage from AI-related mishaps that can impact customer trust and brand value.



Surveyed consumers say they are uncomfortable with how companies use their personal or business information



Surveyed consumers say they have switched brands to protect their privacy

Competitor Differentiation & Superior Value Proposition: Responsibility can set organizations apart from the rest. A US survey from PwC finds that competitive differentiation is the most cited objective for RAI practices, with 46% citing it as a top 3 objective.[33] The same survey reveals that the top benefit from investing in RAI practices is an enhanced customer experience illustrating the case for a superior value proposition.

Comply With Regulation & Avoid Costly Changes: New regulation and potential regulation around genAI is increasing, alongside large fines. The EU AI Act, which is the first major piece of AI regulation globally, has fines whereby noncompliance can cost businesses up to 7% of annual turnover.[34] Companies that proactively implement ethical AI frameworks and governance structures are better positioned to comply with emerging regulations and avoid costly legal battles. With the quickly evolving regulatory landscape, companies that have RAI programs can get ahead and stay ahead, avoiding costly rework or system overhauls. For more on regulation particularly as it relates to American companies, see Box 1.

Mitigate Risk & Drive Sustainable Growth: Proactively addressing ethical concerns minimizes risks and can support greater value generation over time. Companies traditionally focus on a loss aversion strategy, aiming to minimize risks such as regulatory penalties and reputational damage. While this approach addresses immediate concerns, it may overlook opportunities for long-term value creation. A shift towards a value generation perspective is valuable in the genAI space. This involves proactively investing in ethical AI practices, which can support trust, improve customer satisfaction, and drive sustainable growth.[35]

EU AI Act Noncompliance Fines
(up to)

7% *Annual Business Turnover*

Box 1. Policy & Regulatory Considerations for American Companies

The U.S. also has a pre-existing legal and regulatory landscape that applies to AI technologies much like anything else. For example, the U.S. has anti-discrimination laws, intellectual property laws, consumer protection laws, product liability tort, and privacy protections for certain health information. There are also state privacy laws such as California's Consumer Privacy Act (CCPA) and Illinois' Biometric Information Privacy Act (BIPA), which provide further data protection rights to state residents. This paper will not provide a thorough analysis of how existing legal and regulatory landscapes apply to AI technologies, though we note that the Federal Trade Commission has provided guidance on how AI companies should interpret existing consumer protection laws in the context of AI.[36]

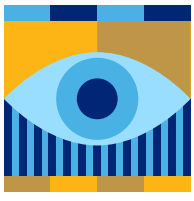
The U.S. does not have an overarching federal law to regulate the development or use of AI, though the National Artificial Intelligence Initiative Act of 2020 (H.R. 6216) was passed into law, which provided funding to numerous U.S. departments and agencies to carry out research, establish standards, and set up advisory groups. Dozens of additional AI bills have been introduced in the years since, but have largely not passed into law. In October 2023, President Joe Biden issued an executive order on the 'Safe, Secure, and Trustworthy Development and Use of AI', which provided guidance across the federal government, and also used existing authorities to require developers of the most powerful AI systems to share their safety test results and other critical information with the government. One year after the publication of the AI EO, the White House reported that more than one hundred tasks across federal agencies had been completed on schedule, though the new administration starting in 2025 brings uncertainty to these advances.[37] Indeed, on the first day of office, President Trump rescinded Biden's 2023 executive order.

Hundreds of state-level AI bills have also been introduced in recent years, and numerous AI bills have been passed. California alone has more than a dozen AI bills that have been signed into law. For example, the California AI Transparency Act requires prominent AI providers to disclose when content has been generated or modified by AI. The Generative AI: Training Data Transparency Act requires genAI developers to publish summaries of the datasets used to develop and train their models; and the Digital Replicas Act helps protect actors and performers from AI-enabled misappropriation of their names, images and likenesses. Other prominent state-level bills include the Colorado AI Act, which requires developers and deployers of high-risk AI systems to take action to prevent algorithmic discrimination, and the Utah AI Policy Act, which requires disclosure of the use of genAI prior to human engagement, and clarifies that companies will be responsible for the statements made by their genAI tools.

Many US AI companies will also be subject to the European Union's (EU) AI Act, which puts requirements on any company providing AI technologies or services to anyone in the EU. These requirements focus on "high risk" AI systems and will require that they are pre-registered in an EU database, and that they are tested and evaluated before being put on the market as well as throughout their lifecycle. The EU AI Act, like other European digital protections, calls for ex-ante regulation, meaning the law is intended to help prevent harms from happening in the first place. U.S. regulation often calls for ex-post oversight instead, which may serve as a deterrent for irresponsible behaviors, but focus on fixing or compensating for harm that has already occurred. Other countries around the world, including China, Canada, and others, have also established AI specific regulations that companies need to take into account.

IV. What are the responsibility risks?

There are various risks and concerns linked to genAI including bias, hallucinations, misinformation, data privacy violations, and more. These can show up in different ways for different daily work use cases.



1. Data Privacy

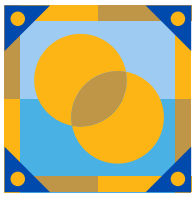
GenAI presents significant data privacy challenges alongside its powerful capabilities. Key concerns include the retention and usage of user data, as models may store interactions to improve future versions, raising questions about long-term privacy and potential copyright infringement. There's a risk of unintended data exposure, where models might generate outputs that reveal personal information or copyrighted material from their training data.[38] LLMs can inadvertently memorize and reproduce sensitive information from their training data. In experiments, a group of researchers were able to extract email addresses, phone numbers, and even credit card numbers that were present in the training data, highlighting the serious privacy implications of these models.[39] The training process itself may infringe on copyrights by reproducing copyrighted works without permission. Additionally, vulnerabilities exist that could allow bad actors to potentially access or misuse private data and copyrighted content.[40] Users can lack clarity on how their information is collected, processed, and stored, and may have limited options to control its usage. This can lead to compliance issues with data protection regulations like GDPR and copyright laws.[41]

To address these concerns, researchers are exploring various measures such as data anonymization, enhanced security protocols, and fair use considerations for AI training. However, balancing innovation with robust privacy protection and copyright compliance remains an ongoing challenge in the rapidly evolving field of genAI.

For example: In July 2023, the Federal Trade Commission (FTC) initiated an investigation into OpenAI, the developer of ChatGPT, focusing on potential consumer harm arising from data privacy issues, violation of consumer protection laws, and the dissemination of false information. The FTC's investigation, which is ongoing, examines whether OpenAI's data security measures are adequate.[42]

Sarah used genAI to analyze customer feedback and generate feature suggestions, but she realized too late that the AI tool had processed sensitive user data without proper anonymization. Concerned about potential privacy violations, she paused the project.

Reflection:
How would you handle this situation? Does your company have guidance in place to inform you?



2. Transparency

As genAI technologies grow more prevalent and powerful, transparency and explainability have become critical concerns. Transparency focuses on openness in the design, development, and deployment of AI systems, ensuring that processes and mechanisms are visible and understandable to stakeholders. This includes sharing information about data sources, model architectures, and decision-making processes. Explainability, on the other hand, involves providing clear, comprehensible reasons or justifications for specific AI decisions or outputs.

The "black box" nature of many AI models, particularly deep learning systems including genAI, makes understanding how decisions are made especially challenging.[43] This opacity poses significant risks in high-stakes domains like healthcare and finance, where issues of accountability, bias, and trust are paramount. Compounding this problem, companies developing genAI systems are often not transparent about the models, withholding details about training data, model architectures, and decision-making processes.[44] This corporate opacity exacerbates the "black box" problem, hindering comprehensive evaluation by users, researchers, and regulators. Currently, there is a lack of transparency from companies about foundation models, particularly related to the training data. The Stanford Foundation Model Openness Index—a measure of openness across 100 metrics—found an average developer score of just 58.[45]

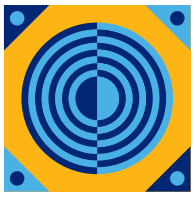
To address these challenges, researchers are pursuing strategies to enhance transparency. Some research is more technical, focusing on understanding the "black" box, while leveraging explainable AI (XAI) techniques aimed to clarify AI decisions in ways that humans can understand.[46] Other research and work is more focused on empowering developers, product managers, and end-users with greater understanding of aspects of the model and its limitations.

For example: Some companies are making efforts to enhance transparency, such as by introducing tools like Model Cards to provide structured summaries of their models' features and limitations. Google has a model card for its model, Gemma, that outlines general model information, the model data, implementation information, ethics and safety considerations, and more. OpenAI released GPT-4, alongside a System Card detailing the model's capabilities, limitations, and the safety measures implemented. The card includes information about the model's architecture, training data, and the steps taken to mitigate risks such as bias, disallowed content, and hallucinations.

Sarah turned to genAI to create a multilingual marketing campaign for a product she was working on, hoping to save time and broaden outreach. But as she reviewed the suggestions, she realized she had no insight into what the model had been trained on, raising concerns about the accuracy, cultural sensitivity, and appropriateness of the messages. Uncomfortable with the risks, Sarah decided to pause and work to see if she could find out more about the model she was using and its training data.

Reflection:

What genAI models do you use in your life and work? Consider exploring documentation about your favorite model (e.g., if it has a Model or System Card).



3. Hallucinations, Inaccuracy

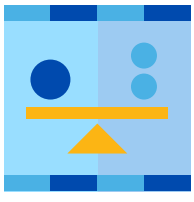
GenAI tools are known to confidently assert false information—or "hallucinate", representing a key barrier to its usefulness and trust. The level of hallucinations can vary for different models, with estimates from prior experiments illustrating ranges from 3–27% depending on the model.[47] Several studies have examined hallucinations in particular domains. For example, a 2024 study by Stanford researchers found that in the legal domain, hallucination rates range from 69–88% in response to specific legal queries amongst top language models. The study found that the models also tend to lack awareness around their errors and reinforce incorrect legal assumptions.[48] Another study examined hallucinations in medical realms. Researchers found that out of 115 references generated by ChatGPT, 47% were fabricated, 46% were authentic but inaccurate, and only 7% were both accurate and authentic.[49]

In addition to being problematic for users with resulting societal impacts, this inaccuracy is a business risk. Inaccuracy is the top risk identified by global executives in 2024, with 63% reporting it as a relevant risk. Yet, only 38% of global executives acknowledge they are working to mitigate inaccuracy.[50]

For example: In a legal case, Mata vs. Avianca, a New York attorney representing a client's injury claim, used ChatGPT to conduct his legal research. The federal judge of the case noticed that the opinion from the attorney had internal citations and quotes that didn't exist and the lawyer was subsequently fined.[51]

Sarah used genAI to draft FAQs for a financial product but noticed it confidently included incorrect details about fees. Alarmed by the inaccuracies, she stopped using it for FAQs and reviewed outputs in greater detail.

Reflection:
How do you review genAI outputs for accuracy?



4. Bias

GenAI tools can exhibit bias in two key ways. First, the tools work better for populations for which they have more training data—and vice versa. This means they currently work worse and have lower productivity benefits for certain populations, including minoritized and vulnerable communities globally.[52] Second, genAI technologies are pattern recognition machines. This means they replicate patterns that exist in society—including harmful or limiting stereotypes, norms, and biases. Various researchers have exposed biases of genAI tools related to gender, race, ethnicity, nationality, language, age, and more.[53] [54] [55] [56] Harms can range from subtle (yet persistent) to more obvious, with variance depending on the use case. Addressing bias can be tricky. In the first case, having more training data for different communities can help to mitigate performance discrepancies. For the second, researchers continue to explore different methods for addressing biases such as finetuning models, using prompt injection to guide outputs, and examining underlying training data. However, this is a tricky problem to solve. Google's AI image generator, Gemini, aimed to promote diversity by depicting historical figures, such as the American Founding Fathers, as people of color. However, this approach led to inaccuracies and public backlash, as it misrepresented historical contexts.[57] Research continues to address this challenge.

For example: GenAI tools asked to output pictures of an American woman have been found to output photos of white women in traditionally American clothing (e.g., jeans, cowboy hats, American flags).[58] If using genAI for writing job recommendations, the technology can more often associate women with feminine adjectives (e.g., warm, kind) versus men with more masculine adjectives (e.g., assertive, decisive).[59] These types of harms may seem subtle, but they can have real implications for people's lives and serve as persistent reminders and implications of limiting and harmful stereotypes.

Sarah used genAI to help design user personas for a new budgeting app but noticed it consistently depicted men as "investors" and women as "budget-conscious shoppers." Concerned about reinforcing stereotypes, she revisited the AI's inputs and processes to ensure the personas were inclusive and avoided stereotyping.

Reflection:
Have you seen outputs in genAI that reflect biases or stereotypes? For what use cases do you need to be mindful of?

5. Safety, Security

GenAI technologies have inherent, unsolved safety and security vulnerabilities and can also be developed and used in ways that threaten people's safety and security. AI safety consists of a combination of technical, human, and systemic factors. It includes investigating the capabilities and limitations of AI systems, under-standing how they interact with humans and are used by people in the real world, and how AI systems are embedded in society, the economy, and the natural environment.[60]

AI safety risks include AI systems that:

1. Fail to perform reliably or effectively under varying conditions, exposing them to errors and failures (e.g. an error in an autonomous vehicle can lead to loss of life);
2. Deceive or subvert human understanding or intentions (e.g. a person engaging with an unsafe AI chatbot might be exposed to self-harm material or convinced to try to harm others); and
3. Are used for censorship, control, and weaponization (e.g. autocratic governments are using genAI to surveil and sway online communications, and nefarious actors are using genAI to lower the barriers to developing cyber weapons.)[61]

GenAI technologies also have security vulnerabilities including susceptibility to prompt injection attacks, where people use malicious inputs to manipulate genAI systems into leaking data or providing dangerous or harmful information that violates their use policy. These vulnerabilities can not simply be patched like traditional software vulnerabilities and are difficult to address. Other key security considerations include training data poisoning and model theft.[62]

For example: Sometimes an LLM can be tricked into providing sensitive or harmful information simply by being asked to adopt the persona of someone who might do such a thing, or if the prompter uses particularly friendly and trusting language. Models can also be tricked by being asked something benign, but being told to ignore the previous prompt and instead provide sensitive information. Indirect prompt injection, where malicious instructions are hidden on a website that the model reads, is another challenge that is hard to guard against.

Sarah used genAI to draft responses for a customer support chatbot, excited about automating common queries. However, during testing, the chatbot unintentionally shared sensitive troubleshooting steps that could expose security vulnerabilities if misused.

Reflection:
Do you ever input sensitive data into different models? How might you better consider data privacy?

At a higher level, concerns include implications for the **future of work, environmental issues**, and **copyright or intellectual property infringement**.

6. Future of Work

GenAI is expected to disrupt and change how work is done, although research varies as to the extent. A 2023 McKinsey study found that, due to genAI, *"automation could take over tasks accounting for 29.5% of hours worked in the US economy by 2030"*.[63] There are four areas where automation could be the highest: customer operations, marketing and sales, software engineering, and product research and development.[64] Meanwhile, researchers at OpenAI found that 80% of the US workforce could have at least 10% of their work tasks affected by the introduction of LLMs with "white collar work" most affected.[65] From the perspective of executives, 43% of respondents in a 2023 survey believe that AI adoption will result in a decrease in organizations' workforce, along with large levels of reskilling.[66]

Automation will not necessarily eliminate jobs, but may result in a reduction of the number of jobs particularly in entry level areas where automation can more easily result in efficiencies. At the same time, genAI can create new jobs. For example, there are new jobs emerging in areas of responsible AI and Chief AI Officers[67]; other areas may have growth potential with increases for worker productivity and job growth (e.g., database analysts), and some jobs will have low potential for exposure with modest growth (e.g., higher education teachers, and personal care workers).[68]

Impacts on labor vary by labor type (i.e. contract versus gig workers) and industry. There have been notable decreases in the gig economy. Researchers found that after the introduction of ChatGPT and image-generating tools, there was "near immediate decreases in posts for online gig workers across job types" including a 21% drop in automation-prone job posts including writing; software, app, and web development; and engineering.[69] From an industry perspective, genAI has indirect impacts on the state of work for certain industries such as writing (e.g., through copyright infringement) and artists (e.g., through the valuation of human creativity).[70]

Finally, there are implications for those in informal work that are developing

AI systems, including AI gig economy workers, AI annotators, and data workers. Click workers are often considered "invisible" and can be exploited labor that informs AI. This work can also have psychological impacts, particularly for annotators that are labeling or interacting with content like hate speech.[71]

7. Environmental Issues

There are environmental concerns in the development of genAI. Machine learning models can have significant carbon footprints due to the production of the needed computing hardware and cloud data center capabilities, the training of the model, and running inference (inferring or predicting outcomes using new input data) with the ML model once it is deployed.[72] Rare earth minerals are integral to the hardware needed to train and power AI (e.g., semiconductors for processing power). Semiconductor factories use large sums of electricity and result in hazardous waste, while also consuming large levels of water.[73] Training models, particularly as models grow increasingly larger, result in large emissions of carbon dioxide and have environmental implications.[74]

While different types of AI require different quantities of computing power to train and run, these are critical considerations to recognize. On the other hand, genAI can be used in ways that help mitigate impacts of climate change and environmental crises, such as allowing managers to better understand energy efficiency.[75] Still, it is important to know that genAI has soaring environmental costs: including estimates that a search driven by genAI uses four to five times the energy of a traditional web search, within years large AI systems "are likely to need as much energy as entire nations", and large amounts of fresh water are needed to cool processors and generate electricity at data centers.[76]

8. Copyright & Intellectual Property Infringement

Copyright and intellectual property infringement are significant concerns in the rapidly evolving field of genAI. These issues primarily arise from two aspects of genAI: the training process and the output generation. The training process of genAI models often involves making digital copies of vast amounts of data, which may include copyrighted works. This raises questions about whether this copying constitutes copyright infringement.[77] The output generated by AI models may also infringe on existing copyrights.

Several lawsuits have been filed against AI companies, alleging copyright infringement in both the training process and the output generation. These cases are still in their early stages, and their outcomes will likely shape the future of copyright law in relation to AI.[78] In December 2023, The New York Times filed a lawsuit against OpenAI and Microsoft for copyright infringement. The Times alleges that these companies used millions of its articles without permission to train their AI models, which now compete with and undermine the Times' content. The lawsuit seeks billions of dollars in damages and aims to prevent OpenAI and Microsoft from using the Times' work in their AI training datasets.[79]

Using genAI responsibly entails proactively considering and addressing potential risks and harms. This does not mean eliminating every risk (as that is not necessarily possible), but taking proactive steps towards mitigating the risk and being transparent about actions taken, as well as limitations.

V. Beware of Challenges to Using genAI Responsibly

There are several challenges to using genAI responsibly—in day-to-day work and in new products.

Organizational & Individual:

1. Lack of *organizational* policies and approaches: Organizations are still grappling with how to respond to growing use of this technology and many lack clear policies governing use of genAI in the workplace, which is related to not understanding what responsibility means or looks like within their organization.[80] Without clear policies and approaches, workers may use genAI haphazardly resulting in potential risks, or may not use it at all, thereby missing out on potential benefits.

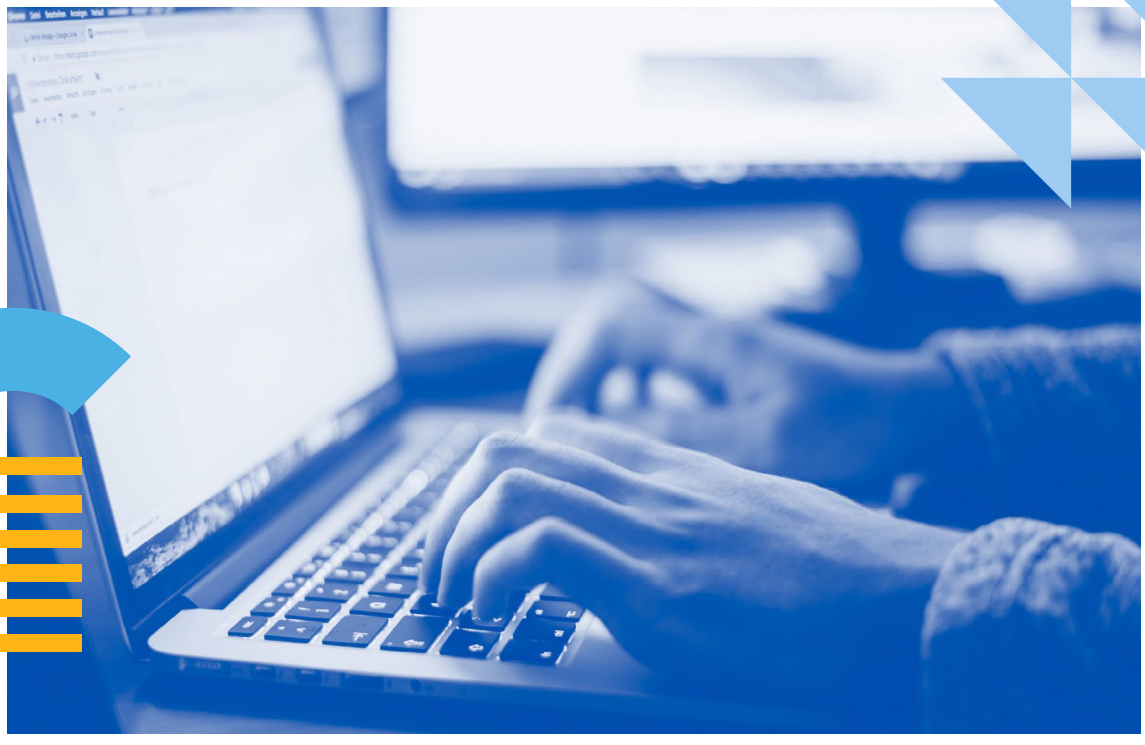
2. Misaligned *organizational* incentives: In using genAI in new products or features, responsible use may require slowing down which can be at odds with speed to market. The tech industry is fast-paced, prizing innovation and disruption. An ethical approach can also sometimes mean blocking or delaying features or products.[81] These tensions can be mitigated by prioritizing long-term product excellence over short-term profits and customer trust and brand value over serious reputational risks and legal consequences. Similarly, organizations prizing productivity may be—even inadvertently—pushing use of genAI without appropriate use training and responsibility guardrails which can result in unforeseen consequences. Taking the effort to build in responsibility, including ensuring that organizational culture supports responsibility, will pay dividends over time.

3. Lack of *individual* education: Individuals using genAI for work or in new products can often be lacking education and training around ethical issues and responsible use of genAI.[82] Many trainings focus on utilizing genAI for reskilling and upskilling, but these don't necessarily include responsibility practices.

4. Lack of *individual* comfort raising issues and considerations on broader org culture: Linked to misaligned incentives, individuals may not be comfortable raising issues especially if it may be at odds with other business priorities or there is a lack of leadership commitment to responsibility. In addition to trainings and support for responsible use of genAI, organizations must have clear policies and approaches, as well as an organizational culture, that support responsibility.[83] [84]

Industry:

5. General immaturity of the field: This is a new technology, which continues to rapidly advance and evolve. Many organizations and managers do not know what responsibility means or looks like, which is reflected in a lack of policies and clear approaches that organizations have for their workers. More broadly, the field does not have general consensus on what responsible use of genAI looks like in regards to the workplace and integrating it into new products. There is a need for consensus building, development of industry-wide standards, as well as greater research and collaborations across organizations and researchers to fill research gaps.



Society:

6. Reinforced inequitable patterns in society: GenAI tools can both exacerbate and reduce existing socioeconomic inequalities. While the tools can support productivity in the workplace, benefits will likely be distributed unevenly.[85] Also, genAI tools—as pattern recognition machines—pick up existing patterns in society (e.g., harmful or limiting stereotypes, inequities), which then become embedded and amplified.[86]

VI. Plays

This section includes ten plays to use genAI responsibly. The first five are for organizational leaders (OL 1-5). The second five are for individual product managers (PM 1-5). The plays are sequential, so it is recommended that they are read and implemented in the order they are presented.

But first, start here:

PLAY 0

Consider whether genAI is the right tool to employ.

Who is Involved:
All employees

About:

While genAI is a technology with many uses and applications, there is a great deal of hype to use it where it might not necessarily add value. GenAI is another tool in the tool chest that can be helpful—or not—for different work use cases or as part of a new product. When considering its use for your organization (as an organizational leader) or for yourself and products you manage (as a PM), ask:

- **Could you accomplish your goals efficiently and effectively without genAI?** If so, consider whether it is even necessary.
- **What is the cost of integrating genAI? How does it scale and is it sustainable for your organization?** Consider economic costs to your organization (e.g., additional work hours, compute), broader societal costs, as well as potential scalability and sustainability over time.
- **Is integrating genAI worth the benefits and repercussions?** This is a place to determine your priorities. Be clear about what you are prioritizing when you make your decision.

Depending on your organization, you may consider gathering a cross-functional team to provide different perspectives and insight in answering the questions.

Then—if the determination is to use genAI in work processes and/or products—proceed to the regular plays.

Here is a table to help you orient to the plays and which ones you should refer to based on your role.

Organizational Leaders (OL)

Whether your organization is using genAI for work use cases, in new products, or both:

→ All plays are relevant

PLAY 1

Ensure leadership recognizes the value of responsible genAI use, develop responsible AI principles, and communicate the organization's commitment to responsibility to all employees.

PLAY 2

Implement policies and accompanying standards to ensure responsible use of generative AI.

PLAY 3

Build a comprehensive responsible AI governance framework that defines key roles, establishes organizational structures, and fosters a culture of shared accountability.

PLAY 4

Update incentives to align performance, product development, and metrics with responsibility.

PLAY 5

Implement tailored training to address gaps and support responsible use of genAI.

Product Managers (PM)

Are you using genAI for work use cases?

→ Go to plays 1 & 5

Are you using genAI in new products or features?

→ All plays are relevant

PLAY 1

Conduct "gut checks" to evaluate responsibility risks in work use cases and product development.

PLAY 2

Choose a model for genAI products by assessing needs and potential risks. Ensure transparency by documenting the model, fine-tuning data, and key considerations.

PLAY 3

Conduct risk assessments and audits for genAI products, involving cross-functional teams, expert oversight, and tools aligned with organizational principles and core risks.

PLAY 4

Implement red-teaming and adversarial testing to uncover vulnerabilities, while capturing and responding to user feedback over time.

PLAY 5

Track your responsibility micro-moments—simple, impactful actions that demonstrate responsible decision-making—and showcase them in performance reviews.

Part A. Organizational Leadership (OL) Plays

OL: PLAY 1

Ensure leadership recognizes the value of responsible genAI use, develop responsible AI principles, and communicate the organization's commitment to responsibility to all employees.

Who is Involved:
C-suite & organizational leaders

About:

Leadership commitment to responsibility is critical. Period. This commitment can be signaled and made clear to employees through communications, as well as explicit responsible AI principles. Responsible AI principles help guide ethical decision making, can inform new strategies and initiatives, impact employee behavior, and result in the adoption of new internal governance approaches such as review processes.[87] AI principles tend to coalesce around 5 topics: bias/ fairness, transparency/ explainability, safety/ security, privacy, and accountability.[88] In the case of genAI, additional principles may include, for example, accuracy and reliability.

These principles have historically been more focused on developing AI tools or products; however principles are also valuable to inform use of genAI. Principles informing the use of AI may be related but slightly different to principles informing the development of AI tools. For example, principles for the use of genAI in organizations may include informed consent (obtaining voluntary and informed agreement from employees to participate in AI-powered interventions), opt-in and easy exits (employees opt-in to use of AI systems and can withdraw without any negative repercussions), communication (informing employees of changes in AI assets and third party relationships), privacy and security (protocols for maintaining privacy and storing employee data), and continual learning.[89] [90]

Business Benefits:

- Informs employee behavior and organizational approaches that mitigate reputational and legal risks over time
- Fosters trust amongst customers and other external stakeholders

How:

While every organization is different and will have different approaches to developing principles, these steps can be helpful.

1. **Ensure leadership is on the same page** regarding commitment to responsibility in developing and using genAI. Then have this commitment be clearly communicated to staff.

2. **Select leaders to lead** the process of principle development.

3. **Review how the organization may use genAI** (including in new products). Conduct consultations with staff across different levels and teams to understand use cases, potential use cases, and concerns.

4. **Examine current and potential benefits and risks from these uses,** while also getting insights and perspectives from external experts. Also meet with legal, IT, and other experts and teams in the organization.

5. **Review principles that exist,** particularly for similar organizations. Identify principles that align to your organizational values and the ways your organization is developing or leveraging genAI. Refine principles with organizational leadership and iterate following staff feedback.

6. **Clearly communicate the principles** and leadership commitment to them, alongside what responsible use of genAI means, to all employees.

Case:

Brookings Institute issued provisional principles regarding adoption of AI in conducting research and other activities. They formed an Emerging Technologies Advisory Group (ETAG) with staff members across every program, business unit, and job level that was chaired by company leadership. ETAG sought to learn more about how genAI is being utilized or could be utilized to inform a set of standards for responsible AI usage matching their institutional values. They gathered information around how the tools were already being used externally, identified where guidance existed (which ranges from banning all uses to allowing it with strong disclosures). The group also conducted a survey to identify where tools were being used across job levels and functions. They identified the most common and riskiest use cases where guidance was necessary and developed four principles to guide use of genAI tools across the organizations: (1) comply with existing Brookings Institution policies; (2) review and validate outputs; (3) protect sensitive data and information; and (4) disclose appropriately. ETAG will continue serving as a resource in developing a comprehensive strategy that considers broader effects of genAI use, such as social impacts and ethical considerations.[91]

Tools & Resources:

- 13 Principles for Using AI Responsibly **(Harvard Business Review, 2023)**
  This article outlines 13 principles for the use of genAI that organizations may consider adopting.

- Responsible AI Principles **(McKinsey, n.d.)**
  This is a list of potential principles that are relevant for the development and use of genAI systems.

OL: PLAY 2

Implement policies and accompanying standards to ensure responsible use of generative AI.

Who is Involved:
C-suite & organizational leaders; legal; privacy; security; trust and safety; responsible AI teams

About:

It is important to have concrete approaches to operationalize AI principles including clear organizational policies and standards that map to different AI principles.

For the use of genAI, an organizational policy that explains acceptable and unacceptable uses of genAI in the workplace and in product design is important. Lack of clarity on both acceptable use and guidance on unacceptable use is common and limits responsible use of genAI.[92] Organizations can identify appropriate use cases (e.g., summarizing meetings) and inappropriate ones (e.g., using proprietary data to inform analyses) that are relevant for their particular workplace and use cases.

If organizations are leveraging genAI in new products, features or services, organizational policy should be explicit on RAI expectations alongside enforcement approaches. For example, policies may mandate risk assessment in the development of products, while also requiring product reviews at the ideation phase, prior to launching, and in an ongoing manner at regular intervals. Product reviews can include ethical and legal reviews, as well as privacy and security reviews—see PM Play 3 for more information. Policies and standards can also inform and are informed by frameworks that can provide more concrete guidance on risk management (see the How section below for more about operationalizing existing frameworks.)

It is important to ensure compliance with the use policies of the foundation model developers themselves. For example, Google has a Generative AI Prohibited Use Policy, stipulating use for its genAI models (e.g., Gemma) that is responsible and legal and outlining how the model may not be used. For example, the model cannot be used for generating and distributing content that can be misinforming or that impersonates an individual without consent. There are similarities and differences between these genAI foundation model use policies. In a meta review of 30 acceptable use policies from companies that develop genAI models, researchers found 127 distinct use restrictions, with notable variance between them.[93]

Business Benefits:

- Mitigates reputational and legal risks over time
- Enhances trust amongst consumers and other external stakeholders

How:

1. **Identify a leader(s) and appropriate partners** to develop the organizational policy/ies. Efforts to implement policies and accompanying standards can be led by the same or similar leaders who led the development of AI policies for the organization. They should coordinate with the organization's RAI principles and be in partnership with legal, IT, and other relevant parties and teams.

2. **Develop AI procurement or licensing guidelines** if your organization procures AI technologies or otherwise enters into a contractual, licensing, or enterprise agreement with a model provider. This can include restrictions on how your organization's data may be stored or used, clarity around responsibility for continuous monitoring, and transparency around updates. These guidelines can ensure alignment with the organization's policies and principles, and may include stipulations about the nature of the model training data, evaluation and risk or impact assessment results, sustainability practices, or other responsible AI practices.

3. **Integrate emerging best practices** and ensure the policies/standards address use of genAI in day-to-day work, as well as use of genAI in new products, features, or services. There are various entities to learn from including national and international standards bodies into their organizational practices and policies. For example, the U.S. National Institute for Standards & Technology (NIST) has published an AI Risk Management Framework and a corresponding Generative AI Profile, which is discussed further in the Tools & resources section of this Play. This profile is important for organizations integrating genAI into new products or features and includes guidelines such as:

    a. Re-assess model risks after fine-tuning or retrieval-augmented generation implementation and for any third-party genAI models deployed for applications and/or use cases that were not evaluated in initial testing.

    b. Implement content filters to prevent the generation of inappropriate, harmful, false, illegal, or violent content related to the particular application.

    c. *Note: NIST and the U.S. AI Safety Institute will continue to update this guidance and provide further resources to organizations to facilitate operationalization, including providing guidance on voluntary reporting templates that can be used to highlight compliance.*

4. **Leverage tools** to support governance processes for genAI models being used.

    a. There are a growing number of governance tools on the market to help organizations manage responsible use of AI, e.g. watsonx from IBM, Saidot, or credo.ai. These tools are useful for organizing different work paths and suggesting checkpoints but should be adapted for your organization and are not a substitute for preliminary and ongoing discussions between organizational stakeholders.

5. **Continue to iterate** the policies over time as lessons are learned and regulation evolves, and update appropriate standards accordingly.

6. **Be transparent to employees** through clear communication about the policies and lessons learned. This is key to reduce information asymmetry that can limit action.

Case:

Salesforce is an interesting case study illustrating how organizations can develop policies for their own internal use of genAI in new products, as well as informing policies of other organizations that leverage their tools. The company both develops and implements AI technologies throughout many of its products and services including its Customer 360 platform and Einstein AI, which it says provides "nearly 200 billion predictions every day across Salesforce's business applications".[94] Organizationally, Salesforce has an Office of Ethical and Humane Use of Technology, which starts with guiding principles inspired by the Universal Declaration of Human Rights.[95] They use the term Ethics by Design to describe the way they translate their principles into daily design, development, and delivery decisions for their products.

One example of this is a module they have developed, available to their community and partners, that helps teach people how to remove bias from data and algorithms. Other policies include Guidelines for Generative AI and an AI Acceptable Use Policy, which applies to customers' use of Salesforce services. Salesforce has also partnered with NIST through the development of the AI Risk Management Framework and as a member of the U.S. AI Safety Institute Consortium.[96]

The University of California at Berkeley has a policy on the Appropriate Use of Generative AI Tools. The policy specifies the following: publicly-available information can be used freely in all genAI tools; the agreements that University of California (UC) has with specific genAI tools, for which the university then allows use with more sensitive information; and prohibited uses of genAI tools, including entering any personal, confidential, proprietary, or otherwise sensitive information into models or prompts, or using the tools for purposes such as grading or disciplinary decision making. The university also calls for abiding by AI developers' usage policies.

At a higher level, UC has several AI principles: Appropriateness; Transparency; Accuracy, Reliability and Safety; Fairness and Non-Discrimination; Privacy and Security; Human Values; Shared Benefit and Prosperity; and Accountability. Building from these, the UC AI Council (which is composed of leaders across the UC system to assist UC's efforts to institutionalize the UC Responsible AI Principles) created a Risk Assessment Guide for AI procurement and administrative purposes. UC Berkeley also has AI risk assessment pre-screening questions that can be used by employees to gauge the level of risk involved for an AI use case (whereby AI is integrated into a product, service or feature at the university). Depending on the level of risk determined, a subcommittee may be engaged and the broader risk assessment conducted. This approach is in the early days and evolving.

Tools & Resources:

For organizations implementing genAI into new products or features:

- Risk Management Framework – Generative AI Profile **(National Institute for Standards & Technology (NIST))** *The NIST GenAI Profile, released in July 2024, is a resource which builds from NIST's AI Risk Management Framework and that provides guidance specific to genAI developers and users. Organizations can use the Profile to inform their own genAI governance and risk management practices by identifying gaps they may have or adopting the overall approach.*[97]

- State of California GenAI Guidelines for Public Sector Procurement, Uses and Training **(GenAI for California)** *The GenAI procurement guidelines released by the California state government March 2024 provide best practices and parameters designed to safely and effectively use AI technologies to improve services for Californians. While it is a tool for the public sector, it remains relevant for private actors.*

- Risk Assessment Guide **(UC AI Council)** *This guide helps assess the risks associated with the procurement, development, and deployment of AI-enabled systems, including data privacy, bias, security, and ethical risks. While it is a tool for universities, it remains relevant for private actors.*

For all organizations where employees are using genAI in day-to-day work:

- AI principles and best practices for employers and worker well-being **(U.S. Department of Labor)** This document outlines principles and good practices that are important to consider when developing workplace policies and standards, particularly regarding the responsible use of genAI in day-to-day work for different employees.[98]

OL: PLAY 3

Build a comprehensive responsible AI governance framework that defines key roles, establishes organizational structures, and fosters a culture of shared accountability.

Who is Involved:
C-suite & organizational leaders

About:

Companies have adopted different organizational structures to support the responsible development and use of AI. There is no one-size-fits-all model, but there are common traits and values among leaders in the space. The most successful organizations strike a balance between clear accountability within roles and shared accountability across the organization.[99] In other words, they have (a) individuals or divisions within the organization that are explicitly tasked with maintaining responsible use of AI as well as (b) a broader culture within the organization that prioritizes responsible use.

Clear accountability with designated responsible AI roles or teams is vital. This can involve having a designated role and/or division within a company that is tasked with maintaining responsible AI. This individual or team can provide training, develop resources, serve as an advisor, and conduct reviews of products prior to and after release. Organizations can use existing divisions within the company (e.g. Trust and Safety) or create new roles and divisions, including through upskilling employees. In many organizations, responsible AI roles were historically separate from product teams. However, over time, organizations have realized the value of incorporating roles with explicit responsibility priorities within product teams.

To complement clear accountability, organizations benefit from structures that provide higher-level oversight and expertise. This may take the form of:

- *Internal AI responsibility councils* address complex ethical challenges and provide strategic guidance, particularly in gray areas.
- *External responsible AI advisory boards* offer additional perspectives, including checks and balances for an organization's practices.

Clear accountability alone is insufficient without fostering a **culture of shared accountability**.[100] Employees across the board need to feel safe bringing safety and ethics concerns to their managers and colleagues without fearing retaliation. Cross-functional collaboration is important across teams and should be informed by shared goals prioritizing safety and responsibility. At a higher level, leadership must actively listen to concerns expressed by employees acting decisively when necessary—such as blocking or delaying shipment of a product that is unsafe or irresponsible according to company principles (see OL Play 1: Clear leadership commitments and OL Play 2: Policies and accompanying standards).

Business Benefits:

- Mitigates reputational and legal risks over time, both proactively and retroactively
- Enhances trust amongst employees, consumers, and other external stakeholders

How:

1. **Identify specific individuals or teams tasked with maintaining responsible genAI use.** This may include appointing a Responsible AI Officer, creating a new team, or creating dedicated roles within existing divisions.

2. **Develop governance frameworks.** This may include establishing internal responsible AI councils and/or external responsible AI advisory boards.

3. **Foster a culture of shared accountability.** Integrate responsibility considerations into shared goals, empower employees to voice concerns, and promote leadership accountability by prioritizing responsibility and safety over speed to market when necessary.

***Case:* These cases highlight the diversity of forms that RAI operationalization can take.**

Large company in tech, explicit AI governance

Microsoft maintains a Responsible AI Office that is accountable to Microsoft's Responsible AI Council, which includes representatives from the company's business, research,
policy, and engineering units. Microsoft's Board of Directors oversees the Council. Within the company there is also an internal AI and ethics committee called AETHER (AI Ethics and Effects in Engineering and Research) that contributes to research and recommendations around responsible AI.

Large company in tech, implicit AI governance

Salesforce's Office of Ethical and Humane Use (OEHU) was formed to promote ethical considerations in the creation and use of tech products. The team considers who products are built for, who they might exclude, and how product use and design can protect vulnerable populations. This office first sat in the Office of Equality at Salesforce, recognizing the crucial links between AI ethics and Diversity, Equity & Inclusion. Later, it spun out of that office to be closer to product teams. Salesforce's culture has always prioritized customer trust, safety, and security as one of the central tenets of company operations and the company's new guidance, research, and governance on genAI continues to fold true to those same tenets. In this way, Salesforce is a great example of a company that has adapted its existing infrastructure to a modern, genAI empowered workforce.

Medium-sized AI first company, integrated AI governance

As an AI-first company, **Anthropic** has fewer AI-specific divisions within its governance structure but principles of Responsible AI are woven into virtually all aspects of its operations. The company has a Responsible Scaling Officer (RSO) that is tasked with approving models and safeguards, reviewing non-compliance reports, and overseeing policy. This office is accountable to Anthropic's Board of Directors. In addition to the RSO, the company maintains a channel by which employees can report non-compliance, including bypassing the RSO and going straight to the Board if the report

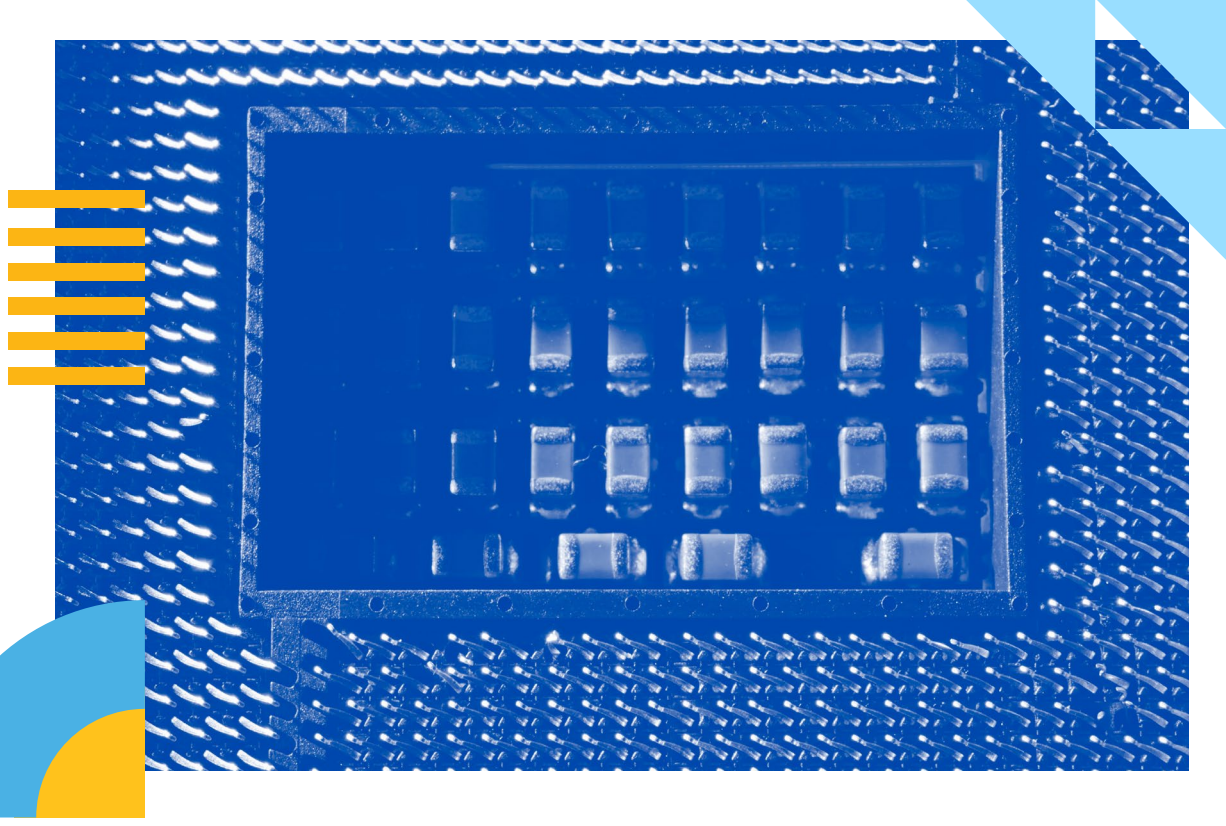
straight to the Board if the report concerns them. Individual teams within the company are tasked with specific functions to maintain Responsible AI, e.g. Frontier Red Team, Trust & Safety, Security and Compliance, Alignment Science.

Large company in financial services, explicit AI governance

Mastercard has taken the form of a two-tier review system to evaluate AI products. The first tier, an AI review board, includes experts from legal, privacy, product, and business. The second tier includes an extensive technical review. By incorporating experts from existing teams, the company can leverage deep institutional knowledge. By having two tiers of review, they can ensure technical evaluation is informed but not dominated by the priorities of other teams.

Tools & Resources:

- **Microsoft's 2024 Responsible AI Transparency Report:** *This report has an overview of Microsoft's organizational structure including its evolution over time and descriptions of several key responsibility-oriented roles and how these roles operate on cross-functional teams.*
- **Anthropic's 2024 Responsible AI Scaling Policy:** *As an AI first company, Anthropic offers a model for how to integrate responsible AI into all aspects of its governance, from top-down roles such as a Chief Scaling Officer to bottom-up policies should as protecting employees who speak publicly about safety concerns.*



OL: PLAY 4

Update incentives to align performance, product development, and metrics with responsibility.

Who is Involved:
C-suite & organizational leaders C-suite & organizational leaders, Human Resources (HR)

About:

Cultivating a culture of responsibility, whereby all employees feel a sense of responsibility in regards to use of emerging technologies and AI, is key for responsible decision-making.[101] This can be done by ensuring employees feel empowered and incentivized to act responsibly. By making responsibility the expectation for employees, leaders support greater shared accountability and ultimately enable more effective responsible decision making day-to-day.

When integrating genAI into new products, incentives for product managers generally focus on speed, efficiency, and shipping products rather than responsible considerations. This results in implementation gaps in regards to responsible use of genAI. By adding incentives that are tied to responsibility, organizational leaders can address these implementation gaps and tensions directly. More specifically, leaders can:

- Ensure product development requirements include responsible AI and map to the organization's responsible AI principles (see OL Play 1).
- Update key performance indicators (KPIs) or objectives and key results (OKRs) to integrate responsibility metrics, redefining what success looks like in development and launch of products.

More broadly for use of genAI across the organization, incentives towards responsibility could be integrated into performance reviews for all employees. Questions in performance reviews can ask about how responsibility is considered in usage of new technologies and actions that employees took. These reviews incentivize responsible decision-making and encourage conversations about challenges and tensions in implementing responsible AI practices. Even in organizations without formal responsibility metrics, product managers and other employees can take the initiative by documenting examples of responsible decision-making in their performance reviews. Doing so not only encourages accountability but also drives broader organizational conversations about responsible AI practices (see PM Play 5).

Business Benefits:

- Mitigates reputational and legal risks over time
- Enhances agency and trust amongst employees

How:

1. **Update individual performance review processes** to include a component around responsible use of technologies. Examples include:
    a. *How did you consider responsibility when using AI or other emerging technologies in different areas of your job?*

b. *What steps did you take to address ethical or safety concerns when using AI or other emerging technologies at work?*

2. **Integrate responsibility into product development requirements.** Review product development requirements and ensure alignment of requirements with the organization's responsible AI principles and policies (see OL Plays 1 & 2). These requirements outline clear steps towards responsibility and accountability.

3. **Revise OKRs / KPIs to reflect responsibility.** OKRs / KPIs can stipulate that certain responsibility metrics are achieved. They might include completion of specific steps outlined in responsible development processes, or achievement of measurable goals linked to responsibility considerations (e.g., transparency, bias mitigation).

4. **Pilot the updated processes to then refine and implement.**

   a. Test the updated performance review processes, product development requirements, and responsibility metrics.

   b. Gather feedback and use it to refine processes and metrics.

   c. Roll out updated practices, while ensuring that leadership champions the effort.

5. **Make sure all employees are aware of how responsibility fits into their role and the expectations.** This is key to a commonplace issue whereby employees feel that responsibility is being taken care of by other teams and are unclear how it fits into their role, which contributes to gaps between AI principles and day-to-day practice.

Case:

Microsoft launched an updated Responsible AI Standard in 2022, which is linked to its six AI principles, which outline product development requirements for responsible AI. Teams developing genAI applications must map, measure, and manage risks throughout the development lifecycle (which aligns with NIST's AI Risk Management Framework). The first step is conducting a responsible AI impact assessment that identifies potential risks and areas to address them. Metrics then measure identified risks for genAI applications and testing occurs to track risk mitigations. Ongoing performance monitoring tracks risks, and there are processes for incident reporting prior to an application being released. Releases are also phased to ensure applications are behaving as expected before being made available to wider audiences.[102]

Tools & Resources:

- Microsoft Responsible AI Standard, v2 – General Requirements: *This document outlines a Responsible AI Standard tracking to responsible AI principles. It includes a set of goals and requirements for AI systems developed by Microsoft.*

- Quick win! Update performance review process & OKRs **(National Institute for Stand UC Berkeley Center for Equity, Gender & Leadership)***: This document, which was developed by one of the authors of this playbook, outlines concrete steps to update review processes & OKRs. Although it is more focused on mitigating bias in AI, it can be adapted for responsible use of genAI.*

OL: PLAY 5

Implement tailored training to address gaps and support responsible use of genAI.

Who is Involved:
Organizational leaders, HR/ Training team

About:

GenAI is a powerful new tool for the workplace, however, how to use it requires education. Analysis from McKinsey finds that with boosts from genAI, up to 27% of current hours worked in Europe and 30% in the US could be automated by 2030. (Without genAI, roughly 20% could still be automated).[103] Reskilling and upskilling will be required as labor markets are impacted and as AI is increasingly integrated into day-to-day work. This includes education around:

- What to use genAI for (i.e. different use cases depending on one's job),
- How to use genAI effectively, and
- How to use it responsibly.

Implementing regular trainings, workshops, and lunch-and-learns around use of genAI is critical to mainstream learnings for responsible use. Sessions should align with the organization's responsible AI principles (see OL Play 1) and leverage resources like this playbook.

For employees developing genAI powered products or features, additional training is important. Our research finds that product managers struggle with "not knowing what they don't know." At the same time, many product managers are eager for education and tools to ensure they embed trust and responsibility into their work.[104] Customized training tied to organizational responsible AI principles, internal practices, and roles can fill these gaps by providing actionable guidance and fostering confidence.

Business Benefits:

- Mitigates reputational and legal risks over time
- Enhances agency and trust amongst employees

How:

1. **Identify training needs and gaps** at your organization as it relates to use of genAI. This can include general employee needs that may be linked to your industry and expectations around genAI usage, as well as specialized needs (such as for employees developing genAI products).

2. **Develop employee-wide training on responsible use in work tasks.**

    a. *How did you consider responsibility when using AI or other emerging technologies in different areas of your job?*

    b. *The structure may include a sharing session around different types of use cases (with breakout groups for people by their role), as well as opportunities for employees to critically reflect on what works well or doesn't and where responsibility concerns lie.*

3. **Develop and implement more specific, deep-dive education** for employees integrating genAI in products or features.

    a. *Consider developing case studies related to your organization so that employees can grapple with relevant challenges and good practices. Ensure tying learning goals to the organization's RAI principles while including any policies or other expectations.*

    b. *Make trainings and workshops interactive, with opportunities for employees to grapple with tradeoffs and ask questions.*

4. **Conduct RAI training with organizational leaders** to ensure they are up-to-date on good RAI practices and understandings.

    a. Organizational leadership training can explore higher-level understandings and outline good practices related to integrating AI ethics into organizations, while presenting the organization's policies and approaches tied to these good practices and the organization's RAI principles.

    b. These trainings could benefit from an outside facilitator depending on the capacity and skills of the organization.

Note: If your organization does not yet have principles and/or internal policies, processes, and tools, initial education can still begin. This can include having external speakers, lunch-and-learns, and discussions. Then, build up to more customized training and education programs that embed good practices within the organization's priorities and structure.

Case:

Google has implemented internal education and trainings on responsible AI for several years. Between 2019 and 2022, Google trained over 32,000 employees on AI Principles, emphasizing the need for ongoing training and dialogue. The Responsible Innovation Challenge has engaged over 13,000 employees through interactive puzzles and quizzes to enhance understanding of ethical concepts. Additionally, Google introduced a two-day Moral Imagination workshop for product teams to explore the ethical implications of AI products, reaching 248 participants from various teams. The company has trainings on AI explainability for internal users and product teams, which was piloted with outside experts. In addition, the company had a more intensive program, its internal AI Principles Ethics Fellows program, which was a six-month fellowship that involved Googlers from 17 global officers. They also have a version for senior leaders in the company.[105]

Tools & Resources:

- **This playbook:** *Use good practices and resources in this playbook, while customizing trainings for your organization.*
- Responsible Generative AI: Online Module **(Microsoft)**: *This is an example module of a 50-minute online training to understand risks and opportunities of genAI systems.*

Part B. Product Manager (PM) Plays

PM : PLAY 1

Conduct "gut checks" to evaluate responsibility risks in work use cases and product development.

Who is Involved:
All employees

About:

As an individual, you have agency to consider usage of genAI for different work use cases. Prior to using a genAI tool for a particular work use case, consider responsibility risks. In some cases, the question is not how to use the gen AI tool, but whether to use it at all for the particular use case.

Similarly, if integrating genAI into a new product or feature, it is important to consider the responsibility risks at the beginning of the product lifecycle by holding a preliminary responsibility discussion. By doing this PMs ensure that conversations about responsibility guide the product's trajectory and inform a more formal risk assessment (see PM Play 3).

Benefits:

- **To you:** Mitigates potential risks that could have repercussions for you and/or your organization.
- **To the organization:** When using genAI in work use cases, reduces unintended consequences and mitigates risk. When using genAI in new products, informs the product's trajectory to mitigate risk, offer a superior value proposition, and enhance trust over time.

How:

When considering gen AI for day-to-day use cases:

1. For each use case, consider the potential risks by asking a series of questions to yourself related to the different responsibility risks: biases, data privacy, transparency, inaccuracy, and safety/security. This does not need to be a formal process, but rather a quick gut check to ensure the usage of genAI in the particular case makes sense and is responsible. Try our Gut Check tool.
2. If using genAI, be transparent about its use by sharing when and how you're using it.

When integrating gen AI into new products:

1. Schedule a meeting to discuss the potential risks and responsibility considerations early in the product development process using the Key questions for PMs when integrating gen AI into new products. Invite responsible AI leads or team members to join. Take notes to document any initial concerns, ideas for mitigation and areas for deeper analysis. These notes can serve as the foundation for a more formal risk assessment if going ahead with the product (see PM Play 3).

2. Schedule regular check-ins as the product continues through the product lifecycle.

3. If using genAI in new products or features, be transparent about its use as well as the responsibility risks and how the team proactively considered and mitigated risks.

Tools & Resources:

- Should I use genAI for this? Take a Gut Check: *This tool is a list of questions for people to ask when considering using genAI for a work use case. Think of it like a responsibility gut check.*
- Key questions for PMs when integrating gen AI into new products

PM : PLAY 2

Choose a model for genAI products by assessing needs and potential risks. Ensure transparency by documenting the model, fine-tuning data, and key considerations.

Who is Involved:
Product managers, responsible or ethical AI leads/teams

About:

Selecting a model to embed within a new product is a critical decision that can significantly impact the product's performance, reliability, and ethical standing. Foundation models vary widely in terms of size, cost, functionality, and inherent risks. Key considerations include the model's customizability, openness (whether it is open-source, proprietary or somewhere on an openness spectrum), data handling policies, biases and limitations. It is essential to assess these factors thoroughly before making a decision. In some cases, your organization is making a new AI model and can build in key considerations from the ground up.

Being transparent about these considerations—not only within your team but also with stakeholders and users—builds trust and accountability. This transparency includes documenting the model selection process, any data used for fine-tuning, and the rationale behind your choices.

Note! Your organization may have guidance on this. Inquire and collaborate with other stakeholders (e.g., the CTO, the responsible technology or AI team).

Benefits:

- **To the Organization:** Reduces reputational and legal risks; provides a superior value proposition that fosters trust; supports compliance
- **To the Customer:** Increases trust and confidence

How:

1. **Assess suitability for your application** by evaluating the capabilities and limitations of potential foundation models. Determine how well each model aligns with your product goals and intended use case. Consider performance metrics such as accuracy, contextual understanding, or language fluency relevant to your application and target users.
2. **Explore potential responsibility risks** associated with the model, including biases, transparency limitations, hallucinations/inaccuracy, and data privacy concerns. Do this by reviewing documentation, model cards or other transparency resources available, and metrics released by the developers. Also review publicly available benchmarks, leaderboards, and independent research to understand the model's performance and limitations.
3. **Define customizability, transparency, and control requirements.**
    a. *More open models:* Offer greater transparency and customization but require technical expertise.

b. *Proprietary models (API or Enterprise):* Provide ease of use and less technical expertise, but are often less transparent, with limited control over and transparency about the training data and model architecture.

c. Understand data usage policies, including how the model processes and stores data.

d. In some cases, your organization may want to build its own generative AI model, which offers the most control, although requires greater resources and expertise.

4. **Develop a risk mitigation plan tailored to your chosen foundation model.** This may include, for example:

   a. Understand and mitigate bias in the foundation model through testing and finetuning.

   b. Ensure the foundation model is able to comply with organizational and legal requirements.

5. **Document the decision making processes.**

   a. Be transparent about the reasons for selecting a certain model by recording criteria used, decision-makers involved, and reasons.

   b. If fine-tuning the model, document data sources, consent and licensing considerations.

   c. Outline the risk mitigation plan taken.

6. **Communicate transparently with stakeholders and users.**

   a. Inform users about the use of genAI in your product and the model used, including capabilities and limitations. Also, provide clear notices where outputs may be uncertain or require human verification.

7. **Establish channels for users to report issues or provide feedback.**

Case:

Adobe launched a family of creative genAI models, **Adobe Firefly**, with an emphasis on transparency and responsible use. Developing their own models had some advantages against other available models such as Stable Diffusion. For example, recognizing the risks associated with copyright infringement that can exist in other popular models, Adobe trained Firefly solely on a dataset composed of public domain content, where the copyright has expired, and licensed content, such as Adobe Stock, with contributors compensated for its use. By doing so, it mitigated legal risks related to intellectual property. There was still pushback from some Adobe Stock creators, who felt that it was unethical for Adobe to train Firefly on their IP and then flood Adobe Stock with AI-generated images, which can reduce revenue for the real artists. However, Adobe contends they have been transparent about the data sources and the limitations of Firefly. It provided detailed documentation on how the model was trained and offered disclaimers about potential biases or inaccuracies. Adobe also states that it does not train Firefly on its users' data, respects artists' rights, and addresses data privacy concerns.

Tools & Resources:

- The Foundation Model Transparency Index **(Stanford):** *Outlines how transparent different models are and includes transparency reports by a variety of model developers.*
- Responsible AI Tools and Practices **(Microsoft)**: *Includes resources for AI impact assessments that can be leveraged.*

PM : PLAY 3

Conduct risk assessments and audits for genAI products, involving cross-functional teams, expert oversight, and tools aligned with organizational principles and core risks.

Who is Involved:
Product managers, responsible or ethical AI leads/teams

About:

Before and after deploying a generative AI (genAI) product or feature, conducting thorough reviews and audits is essential. These evaluations should address key risks associated with genAI: data privacy, transparency/explainability, hallucinations/ inaccuracy, bias, and safety/security.

While evaluations or reviews are not necessarily done by product teams, it is important for product managers and teams to ensure they are completed, work with collaborators throughout the process, and incorporate results to strengthen the product.

Benefits:

- **To the Organization:** Identifies and mitigates potential legal, ethical, and safety risks early on, protecting brand integrity and ensuring regulatory compliance
- **To the Customer:** Builds trust by ensuring products are safe, accurate, and respectful of user rights and data privacy

How:

1. **Conduct an initial risk assessment** collaborating with responsible AI experts and involve cross-functional teams (legal, ethical, technical).
    a. Building from the initial discussion (in PM Play 1), conduct an evaluation to assess potential harms and responsibility risks of the product. Categories for the assessment can be informed by AI principles at the organization and may relate to bias, privacy, transparency, inaccuracy, safety, and security.
    b. Based on the risk assessment, determine if the product aligns with organizational principles and legal compliance.
    c. Categorize the levels of risk across responsibility areas to inform ongoing management and mitigation plans.
2. **Establish a review framework** informed by the risk assessment.
    a. Set clear objectives, evaluation criteria and benchmarks working with responsible AI experts and cross-functional stakeholders (e..g, legal counsel, data scientists).
    b. Identify roles for different stakeholders.

3. **Conduct pre-deployment reviews.** Test the product across responsibility risks across a variety of scenarios. Different tools can be leveraged in different responsibility areas, such as:

    a. For security and safety: Apply security assessment tools to identify vulnerabilities like susceptibility to prompt injection attacks and data poisoning.

    b. For explainability: Use interpretability tools to understand how the model makes decisions.

    c. For bias: Software can help detect and quantify certain biases in model outputs. This is not exhaustive, however, and teams should also engage domain experts and social scientists to understand potential biases and mitigation approaches.

    d. For inaccuracy: Test outputs rigorously, particularly for high-stakes applications, and create workflows for human oversight when necessary.

4. **Monitor post-deployment performance.**

    a. Regularly evaluate the product and set up monitoring systems to track key metrics, while also doing periodic audits.

    b. Ensure that there are clear user feedback channels where users can report issues or provide feedback on AI interactions.

5. **Mitigate risks and document actions.**

    a. Address risks promptly and maintain transparency with different stakeholders.

    b. Document risks identified, decisions made, and steps taken.

    c. Share findings and mitigations with stakeholders.

6. **Engage external expertise.** Work with domain experts and social scientists who can provide relevant insights. For high-risk applications, consider external evaluations to validate compliance and mitigate bias or other risks.

Case:

Microsoft has implemented a rigorous AI ethics review process for all AI products and features. Before deployment, products undergo an audit that assesses compliance with Microsoft's Responsible AI Principles, which include fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability. For instance, when developing their AI-powered chatbot in Microsoft Teams, the product team collaborated with the Office of Responsible AI and the Responsible AI Ethics and Effects in Engineering and Research (Aether) Committee to identify potential risks such as data privacy concerns and the potential for generating inappropriate content. They used internal tools to evaluate the model's outputs for biases and other ethical considerations.

Salesforce leverages an "Ethics by Design" process that embeds ethical considerations throughout the product development lifecycle. They conduct regular reviews of their AI tool, like EinsteinAI, ensuring they meet various ethical standards and avoid causing unintentional harm. Salesforce provides the guidelines they use to assess models for fairness, leveraging both internal

resources and external tools, like the Consequence Scanning methodology. They document their findings and make necessary adjustments before AI product releases.

Google has an Equitable AI Research Roundtable (EARR) with experts and social scientists across different domains that product teams can speak with to get feedback on responsibility considerations for different products. This comes from a recognition that software tools are limited in understanding the nuances of considerations such as bias, and help provide more holistic product reviews.

Tools & Resources:

- AI Risk Management Framework **(NIST):** *This framework provides a customizable approach to mapping, measuring, managing, and governing risks that may emerge throughout the AI lifecycle.*
- Generative AI Profile **(NIST)**: *This profile provides additional risk management guidance tailored to genAI specifically.*
- Risk Management Profile for Artificial Intelligence and Human Rights **(NIST):** *This profile provides additional risk management guidance focused on human rights.*
- Fundamental Rights Impact Assessment **(Aligner)**: *This tool helps organizations conduct an AI fundamental rights impact assessment, a new requirement in some cases under the EU AI Act.*
- Responsible AI Impact Assessment Template **(Microsoft)**: *This template can be used by organizations to assess the impacts of their AI tools and responsible AI practices.*

PM : PLAY 4

Implement red-teaming and adversarial testing to uncover vulnerabilities, while capturing and responding to user feedback over time.

Who is Involved:
Product managers, responsible or ethical AI leads/teams, privacy and security teams, trust and safety teams, external evaluators

About:

When deploying genAI in new products or features, it's crucial to anticipate and mitigate potential security and safety risks by thoroughly testing the system before and after launch. Red-teaming and adversarial testing involve intentionally challenging the AI model and simulating attacks or misuse cases to identify vulnerabilities, biases, or undesirable behaviors that could harm users or compromise the system's integrity. This helps stress-test genAI models to identify and mitigate security and safety risks[106] including technical vulnerabilities.[107] Capturing and responding to user feedback post-deployment complements these practices, ensuring continuous improvement and sustained user trust.

While red-teaming and adversarial testing are not necessarily done by product teams, it is important for product managers and teams to ensure they are completed, work with collaborators throughout the process, and incorporate results to strengthen the product.

Benefits:

- **To the Organization:** Strengthens security and resilience by uncovering vulnerabilities; reduces reputational, legal, and financial risks
- **To the Customer:** Enhances confidence and trust

How:

1. **Assemble a team with diverse backgrounds and areas of expertise**, ensuring that your testing can cover a wide range of people and real-world scenarios. Consider involving external experts or third-party teams to provide objective evaluations, especially for high-risk applications.

2. **Try to "break" the AI system**, by inputting malicious, unexpected, or ambiguous data to see how it responds. The aim is to uncover situations where the AI might fail, generate biased or harmful content, or be manipulated into doing something it's not supposed to.

    a. Use techniques such as prompt injections to manipulate responses, queries to elicit biased or harmful outputs, and attempts to bypass safeguards.

3. **Establish a standardized protocol** for the particular product or feature to keep efforts consistent. The protocol can outline methods, tools, and criteria. It can include:

    a. Predefined testing goals.

    b. A structured approach that can be repeated, ensuring that all potential risks are thoroughly evaluated across different product versions and updates.

4. **Document any vulnerabilities or issues discovered in testing.**

    a. Documentation should include detailed descriptions of the problems, the context in which they occurred, and the potential impact on users and the organization.

    b. Use these insights to refine the AI model, adjust the training data, or add safeguards to address any identified weaknesses.

5. **Mitigate risks and document actions.**

    a. Address risks promptly and maintain transparency with different stakeholders.

    b. Document risks identified, decisions made, and steps taken.

    c. Share findings and mitigations with stakeholders.

6. **Set up user feedback mechanisms** (e.g., in-app feedback forms, dedicated customer support lines, or community forums) and regularly monitor and address feedback.

Example: Before releasing Bard, Google recruited hundreds of Googlers with various backgrounds (demographic, professional) to intentionally violate the use policies and test the service. Google continues to conduct internal adversarial tests to inform expansions and future releases. In addition to this internal work, Google seeks input from communities to understand societal contexts early on to inform stress testing. For example, they partnered with MLCommons and Kaggle to create Adversarial Nibbler," a public AI competition to crowdsource adversarial prompts to stress-test text-to-image models, with the goal of identifying unseen gaps, or "unknown unknowns," in how image generation models are evaluated." The company evolves its red-teaming efforts over time to inform new security frameworks and engages the public in red-teaming such as at conferences (Google, 2024).

Tools & Resources:

- ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) **(MITRE):** *Provides an overview of adversary tactics and techniques against AI-enabled systems based on real-world attack observations and realistic demonstrations from AI red teams and security groups.*
- AI Chat Model Benchmark **(ML Commons):** *Evaluates AI chat models for physical, non-physical, and contextual hazards.*
- NIST resources *on AI red-teaming including their open-source software, Dioptra, which can support testing and red-teaming of genAI models.*
- Vivaria **(MITRE):** *An open-source platform for running genAI testing and evaluations.*
- Generative AI Red Teaming Challenge: Transparency Report **(Microsoft)**: *This report highlights key insights from a major AI red teaming challenge.*

PM : PLAY 5

Track your responsibility micro-moments—simple, impactful actions that demonstrate responsible decision-making—and showcase them in performance reviews.

Who is Involved:
All employees

About:

Every employee has the individual agency to act upon responsibility micro-moments—which are small, simple decisions towards responsibility when using genAI. These micro-moments can be found throughout the plays mentioned in this playbook. They can range from quick gut checks around when and how to use a genAI tool, to participating in lunch-and-learns or trainings on responsible AI, to conducting assessments incorporating responsibility when deciding which type of model to leverage for a new product. Employees can then track and reference these actions in their own performance reviews—such as in metrics related to responsibility, ethics, personal initiative, or social impact.

This is important because top-down incentives for responsibility in regards to genAI use may not always exist, but product managers (and employees more broadly) can still exercise their personal agency to take advantage of responsibility micro-moments. Tracking these moments and referencing them in performance reviews in ways that align with their organization's values and principles can look favorable for employees who can be seen as leaders in regards to ethical use of this emerging technology.

Throughout this process, it is important for employees to continue approaching the technology with curiosity and healthy skepticism while feeling empowered to take those responsibility micro-moments. Employees can continue having conversations with others who are using the technology to chat through concerns and discuss challenges or opportunities, and continue learning through trainings and workshops on responsibility and ethical concerns. These actions help employees to get ahead and stay ahead.

Benefits:

- **To you:** Sets you apart as an employee who embodies the organization's values and brings to life its principles
- **To the Organization:** Addresses the gap that often exists between an organization's values and AI principles and day-to-day actions by employees; mitigates risks over time

How:

1. **Identify plays relevant for you and take those responsibility actions,** which can range from small decisions like gut checks and decisions around which models to use through a responsibility lens; to larger actions around assessments and adversarial testing in new genAI products.
2. **Track actions taken and reference them in performance reviews** that align with your organization's values and AI principles (if they exist).
3. **Share about actions taken** with colleagues and continue having discussions with colleagues to encourage them to exercise their own personal agency for responsibility.

Tools:

- **This playbook!** *Use this playbook and exercise your own agency to identify day-to-day actions towards responsible use of genAI.*

Call to Action

Sarah, initially unsure about using genAI responsibly, turned to this playbook for guidance. Applying the plays, she grew confident in using it in both her daily work and a new product feature: an AI-powered content generator that created engaging financial tips and summaries for users. After integrating responsibility considerations from the start, Sarah earned customer trust, boosting adoption and proving the value of responsible innovation and product management.

By implementing approaches for responsible use of genAI, business leaders and product managers like Sarah can unlock the immense potential of genAI for their products and organizations, while supporting positive societal impact more broadly. They can also mitigate responsibility risks and lead in this new technology era.

Our goal with this Playbook is to guide you towards concrete and meaningful actions, which are grounded in research and evidence. It is now up to you to sustainably capitalize on the transformative potential of genAI.

Appendix

This Playbook was authored by Genevieve Smith (UC Berkeley), Brian Lattimore (Stanford University), Natalia Luka (UC Berkeley), and Jessica Newman (UC Berkeley); with research support and guidance from Merrick Osborne (UC Berkeley), Brandie Nonnecke (UC Berkeley) and Brent Mittelstadt (University of Oxford). The Playbook drew from a research project conducted by the aforementioned authors and researchers that included 25 interviewees and 300 survey respondents (anonymized as this research informed an academic paper). We greatly appreciate the insights shared with us via the interviews and survey. The playbook benefited from the helpful feedback provided in prototyping sessions from: Connor Sullivan (Google), David Parham (Fiutur), Dominique Wimmer (Ex-Google, Ex-Meta), Hamsa Pillai (Airvue), John Reed (Splunk), Mark Weeks (ScOp VC), Michael Boone (Nvidia), Neha Surendranath (MHK), Parisa Assar (Intuit), RK Neelakanadan (Google), and Teginder Singh (Google); as well as contributions from Ariana Haider (UC Berkeley) and Nadia Abbasi (UC Berkeley). Invaluable support was provided by Reena Jana (Google). The design was conducted by Ellen O'Reilly.

The Playbook was made possible through funding provided by Google.

Tool 1. Should I use genAI for this? Take a Gut Check.

About:
This tool is a list of questions for people to ask when considering using genAI for a work use case. Think of it like a responsibility gut check.

How to Use:
Review the questions and follow the decision tree.

---

1. Bias: ***Could the AI outputs reflect or reinforce certain stereotypes about certain populations?***

a. **NO:** Proceed to the next question.

b. **YES:**

i. Critically review the outputs and reassess the prompts or data.

ii. Consider not using the tool for this case.

2. Hallucinations / Inaccuracy: ***Does the use case tolerate occasional inaccuracies or errors?***

a. **YES:** Proceed to the next question.

b. **NO:**

i. Validate the AI's reliability or introduce human review before use.

ii. Consider not using the tool for this use case.

3. Data Privacy Violations: ***Does the use case involve inputting sensitive or proprietary data?***

a. **NO:** Proceed to the next question.

b. **YES:** Confirm encryption and privacy measures are in place. Avoid using AI if these cannot be ensured.

4. Lack of Transparency: ***Can the AI's decision-making process be explained and justified?***

a. **YES:** Proceed to the next question..

b. **NO:** Avoid using AI for decisions that require accountability or user trust.

5. Safety and Security: ***Could the AI outputs harm users or be exploited maliciously?***

a. **NO:** The use case may be suitable for AI.

b. **YES:**

i. Add safeguards to prevent misuse.

ii. Consider not using the tool for this case.

Tool 2. Key questions for PMs when integrating genAI into new products

About:
This tool is for PMs that are integrating genAI into a new product or feature. It includes a list of key questions for each stage of the product life cycle.

How to Use:

A. Schedule a meeting to discuss the potential risks and responsibility considerations early in the product development process. Invite responsible AI leads or team members to join. Take notes in the document to save and refer to.

B. Schedule regular minute check-ins as the product continues through the product lifecycle.

Why Use This:
By addressing these areas, the PM can proactively mitigate challenges, align the product with regulatory and user expectations, and build a responsible and trustworthy genAI-powered product.

1. Bias

How might this product or feature reflect biases or stereotypes? Will this product or feature work better for certain groups?

How might we ensure that the product or feature empowers all users and works equally well for different populations, paying particular attention to marginalized people?

Consider the training data, testing scenarios, and the diversity of the end-users the product will serve.

2. Hallucinations / Inaccuracy

How will we validate the accuracy and reliability of the AI outputs, especially in scenarios where errors could harm users or damage trust?

This includes building mechanisms for error detection, human oversight, and clear disclaimers for uncertain outputs.

3. Data Privacy

What measures will we implement to protect user and proprietary data from breaches or misuse, and to ensure compliance with privacy regulations?

Think about how data is stored, processed, and handled during training, fine-tuning, and in live environments.

4. Transparency

What are the transparency needs of different stakeholders (e.g., users, business leaders, other teams)?

How will we explain the role of AI in the product to stakeholders and users, ensuring they understand how and why the AI generates its outputs?

Transparency is essential for accountability, user trust, and regulatory compliance; it also looks different to different people so may need different approaches.

5. Safety and Security

What safeguards are in place to prevent the misuse of the AI, ensure its outputs are safe, and protect the system from malicious attacks?

This includes addressing vulnerabilities, misuse scenarios, and user safety in live environments.

6. Environmental Issues

What may be some environmental considerations or costs of using genAI for this product **initially** and **over time**?

As a PM, you may not know the full scale of environmental considerations or costs. Regardless, it is important to keep in mind.

You can always say No.

If genAI is not the right solution to your problem or cannot be used responsibly in its current form, the answer may be not to use it.

Tool 3. Key questions for PMs when using genAI for work use cases across the product lifecycle

About:

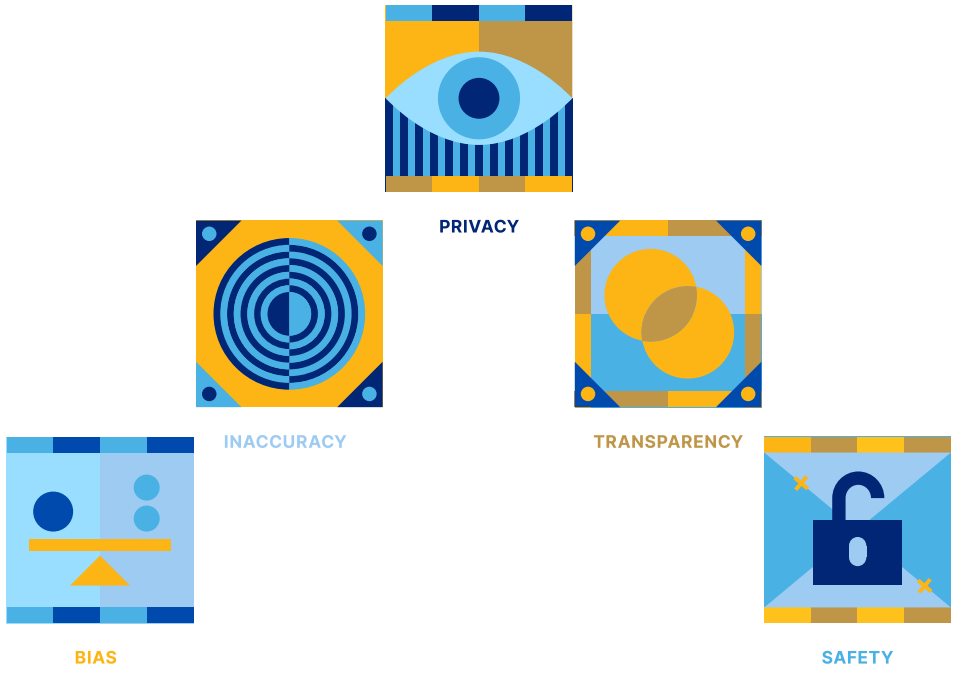
This tool includes genAI use case examples across the product lifecycle for product managers to consider. It also includes key responsibility questions to ask when using genAI at different stages of the product lifecycle.

How to Use:

Meet with your team to explore different use cases of genAI that you may want to use through the product lifecycle. Use the key questions to have a discussion to support responsibility among them team when using genAI throughout the product lifecycle.

	Examples of genAI Use	***Key Questions to Ask***
Brainstorm / Ideation	• Brainstorm product ideas • Market research • Generate product names & branding concepts	***Bias:*** Are we relying on diverse datasets and inputs to ensure inclusive idea generation? ***Inaccuracy:*** How do we verify the quality and feasibility of AI-suggested ideas? ***Privacy:*** Are customer or proprietary insights being anonymized during analysis? ***Transparency:*** Can we trace the source of AI-generated ideas and justify their inclusion? ***Safety:*** How do we prevent AI misuse during early-stage experimentation?
Define	• Generate drafts of product requirement documents or user stories • Summarize complex requirements for stakeholders • Analyze customer pain points and prioritize features	***Bias:*** Are AI-derived requirements reflective of a wide range of users and their lived realities? ***Inaccuracy:*** How do we validate that AI-generated specifications align with actual customer expectations? ***Privacy:*** Are we protecting sensitive data during AI-powered requirement drafting or prioritization? ***Transparency:*** Can we document how AI insights influenced requirement prioritization? ***Safety:*** How do we ensure that sensitive design decisions are securely handled by AI tools?
Design	• Prototyping with wireframes, mockups, or user interface (UI) designs • Generate user experience (UX) recommendations • Suggest personalization of features and design elements based on audience segments and preferences	***Bias:*** Are AI-generated designs or suggestions accessible and inclusive for all user groups? ***Inaccuracy:*** How do we confirm the functionality and usability of AI-proposed designs? ***Privacy:*** Are proprietary design concepts securely managed when using AI tools? ***Transparency:*** Can we explain and justify AI-driven design choices to stakeholders? ***Safety:*** How do we prevent design outputs from inadvertently introducing vulnerabilities?

	Examples of genAI Use	*Key Questions to Ask*
Test	• Create test cases and scenarios for quality assurance (QA) • Simulate user interactions and predict potential usability issues	***Bias:*** Are AI-generated test cases accounting for diverse user contexts and edge cases? ***Inaccuracy:*** How do we validate that AI-identified bugs or insights are correct and actionable? ***Privacy:*** Are test environments secured to prevent data exposure during AI analysis? ***Transparency:*** Can we explain how AI tools contributed to test outcomes or recommendations? ***Safety:*** What safeguards are in place to protect testing environments from AI-related risks?
Launch / ongoing management	• Generate marketing copy, FAQs, user guides, and other launch materials • Customer support assistants • Performance analysis	***Bias:*** Are AI-driven decisions (e.g., chatbots or recommendations) fair and unbiased across all user segments? ***Inaccuracy:*** How do we monitor and address inaccurate or harmful outputs from AI in live environments? ***Privacy:*** How do we protect customer data collected or processed by AI during operations? ***Transparency:*** Are end-users informed about the AI's role in delivering features or services? ***Safety:*** How do we safeguard against security breaches or misuse of AI-powered features?



How the playbook & plays tie to research

This playbook built from a literature review on research in AI ethics in organizations, as well as an academic study exploring responsible use of genAI by product managers. The academic study—drawing on a theoretical framework grounded in Institutional theory coupled with Agency Theory and the concept of Diffusion of Responsibility—employed a mixed methods analytical framework, drawing on 25 interviews with PMs and a global survey with 300 respondents in product management related roles. There are some clear ways the findings connect to the plays in this playbook.

This research informed the Organizational Leadership plays in several ways. First, AI principles that are tied to organizational values and clear leadership commitment that is then communicated to all staff help address the normative pillar (social norms, shared values, expectations) and cultural-cognitive pillar (deeply embedded beliefs) (from Institutional Theory) that are important for institutional processes. Standards and guidelines tied to principles, as well as trainings and resources for all PMs, similarly help build the normative and cultural-cognitive pillars. Clarifying roles and expectations for responsibility are also important to minimize moral hazard, information asymmetry and "diffusion of responsibility" (which is when perception of shared accountability among teams and lack of clarity on role expectations exacerbates uncertainty and inaction). The lack of incentives tied to AI principles leads to a tension between organizational leaders and their commitments and product managers (related to a concept of principal-agent tension in Agency Theory). This can be mitigated through implementation of incentives tied to company values and principles that are connected to day-to-day action.

The research informed the Product Manager plays in several ways as well. Product managers who take advantage of education, training and resources support the normative and cultural-cognitive pillar (Institutional Theory). The other recommendations and micro-moments also support the cultural cognitive pillar and fill a key gap: Organizational leaders are lagging in tying incentives to AI principles, which is partially linked to the priority of speed-to-market and the hype in the industry overall. There is thus an important role for PMs to play in exercising their own agency and taking advantage of responsibility "micro-moments". These bottom-up actions are key in an industry that is obsessed with speed and fraught with the tension that the priority of speed presents. Implementing these responsibility "micro-moments" allows for a "recoupling" to occur between organizational values / AI principles and day-to-day action. PMs can also create bottom-up incentives through discussing their responsibility micro-moments in performance reviews, tying them to the company values and principles in a way that their supervisors can recognize and value.

At a higher level, there remains a need for industry-wide standards, including for enhanced transparency to minimize uncertainty and opacity that is common in the tech industry. The immaturity and uncertainty in the industry contributes to widespread uncertainty about what responsibility means and looks like. Regulation also plays an important role to minimize uncertainty.

Read the academic paper, **"Responsible Generative AI Use by Product Managers: Recoupling Ethical Principles and Practices."**

Endnotes

1. This executive summary was developed with the assistance of NotebookLM.
2. Sarah is fictional, but based on real people and ways that product managers are using genAI (building from our interviews with product managers in a range of industries)
3. Kenthapadi, K., Lakkaraju, H., Rajani, H. (2023). Generative AI Tutorial from ICML. Retrieved from link.
4. Eiras et al. (2024). Near to Mid-term Risks and Opportunities of Open-Source Generative AI. Retrieved from link.
5. AWS. (2024). What are foundation models? Retrieved from link.
6. Computer & Computing Industry Association. (n.d.) AI foundation models explained. Retrieved from link.
7. Bick, A., Blandin, A. & Deming, D. (2024). The rapid adoption of generative AI. Retrieved from link.
8. McKinsey. (2024). The state of generative AI in early 2024. Retrieved from link.
9. Bick, A., Blandin, A. & Deming, D. (2024). The rapid adoption of generative AI. Retrieved from link.
10. Bick, A., Blandin, A. & Deming, D. (2024). The rapid adoption of generative AI. Retrieved from link.
11. The Project on Workforce. (2024). The rapid adoption of generative AI. Harvard Kennedy School Retrieved from link.
12. McKinsey. (2024). The state of generative AI in early 2024. Retrieved from link.
13. ChatGPT Enterprise webpage: link.
14. McKinsey. (2024). The state of generative AI in early 2024. Retrieved from link.
15. [4] https://www.canva.com/magic-write
16. https://www.canva.com/help/using-magic-media
17. https://newsroom.spotify.com/2023-02-22/spotify-debuts-a-new-ai-dj-right-in-your-pocket
18. McKinsey. (2024). The state of generative AI in early 2024. Retrieved from link.
19. PwC. (2024). Tech translated: Foundation models. Retrieved from link.
20. McKinsey. (2023). The economic potential of generatie AI. link.
21. Dell'Acqua et al. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Harvard Business School. Retrieved from link.
22. Bick, A., Blandin, A. & Deming, D. (2024). The rapid adoption of generative AI. Retrieved from link.
23. McKinsey. (2024). The state of generative AI in early 2024. Link
24. McKinsey. (2024). The state of generative AI in early 2024. Link
25. Tilley, A. & Kruppa, M. (2024). Apple Restricts Employee Use of ChatGPT, Joining Other Companies Wary of Leaks. Wall Street Journal. Link
26. List compiled from Business Insider. Link
27. (2024). OpenAI and Apple announce partnership to integrate ChatGPT into Apple experiences. OpenAI. Link
28. British Broadcasting Corporation. (2024). *Airline held liable for its chatbot giving passenger bad advice – what this means for travellers*. BBC Travel. Link
29. McKinsey. (2024). The state of generative AI in early 2024. Link
30. Reuel et al. (2024). Responsible AI in the Global Context: Maturity Model and Survey. Link
31. IBM Institute for Business Value. (2023). The CEO's guide to generative AI: Responsible AI & ethics. IBM. Link
32. Renieris, E. M., Kiron, D., & Mills, S. (2024). Artificial intelligence disclosures are key to customer trust. MIT Sloan Management Review. Link
33. PwC. (2024). 2024 US Responsible AI Survey. Link
34. EU Artificial Intelligence Act. (2024). Article 99: Penalties. Link
35. Author(s). (2024). On the ROI of AI ethics and governance investments: From loss aversion to value generation. California Management Review. Link
36. Atleson, M. (2023). The luring test: AI and the engineering of consumer trust. Federal Trade Commission. Link
37. The White House. (2024, October 30). Fact sheet: Key AI accomplishments in the year since the Biden-Harris administration's landmark executive order. Link
38. Sobel, B. (2023). Copyright accelerationism. Link
39. Carlin et al. (2021). Extracting training data from large language models. Proceedings of the 30th USENIX Security Symposium. August 11–13, 2021.

40. Levendowski, (2018). How copyright law can fix artificial intelligence's implicit bias problem. Washington Law Review, 2(93).

41. Confederation of European Data Protection Organisations. (2023). Generative AI: The data protection implications. Link

42. Zakrzewski, C. (2023). FTC investigates OpenAI over data leak and ChatGPT's inaccuracy. *Washington Post*. Link

43. von Eschenbach, W. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology*, 34: 1607–1622. Link

44. Lemley, M. . (2024). How Generative AI Turns Copyright Law Upside Down. *Science and Technology Law Review, 25*(2). Link

45. Foundation model transparency index. Stanford. Accessed on November 25, 2024. Link

46. Sobel, B. (2023). Copyright accelerationism. Link

47. Hughes, S. (2023, November 6). *Cut the bull...Detecting hallucinations in large language models*. Vectara. Link

48. Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Hallucinating law: *Legal mistakes with large language models are pervasive*. Stanford HAI. Link

49. Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., & Miller, L. E. (2023). High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus*, 15(5), e39238. Link

50. McKinsey. (2024). The state of generative AI in early 2024. Link

51. Feuer, A. (2023, May 27). A lawyer's filing in a case against Avianca was full of fake citations. *The New York Times*. Link

52. Fleisig, E., Smith, G., Bossi, M., Rustagi, I., Yin, X., & Klein, D. (2024). Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. Link

53. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023, November). *Demographic stereotypes in text-to-image generation*. Stanford HAI. Link

54. Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). Bias in generative AI. Link

55. Fleisig, E., Smith, G., Bossi, M., Rustagi, I., Yin, X., & Klein, D. (2024). Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. Link

56. Ghosh, S. & Caliskan, A. (2023). ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. Upcoming Publication in AAAI/ACM Conference on AI, Ethics, and Society 2023.

57. BBC News. (2024). *Google AI diversity push sparks backlash*. BBC. Link

58. Turk, V. (2023, October 10). How AI reduces the world to stereotypes. *Rest of World*. Link

59. Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., & Peng, N. (2023). Kelly is a warm person, Joseph is a role model: Gender biases in LLM-generated reference letters. *arXiv*. Link

60. Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., & Isaac, W. (2023). Sociotechnical safety evaluation of generative AI systems. *arXiv*. Link

61. Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2024). *The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence*. MIT FutureTech. Link

62. OWASP Foundation. (2023). *OWASP Top 10 for large language model applications – 2023* (Version 1.1). Link

63. Ellingrud, K., Sanghvi, S., Dandona, G. S., Madgavkar, A., Chui, M., White, O., & Hasebe, P. (2023, July 26). *Generative AI and the future of work in America*. McKinsey Global Institute. Link

64. Yee, L., & Chui, M. (2023). The economic potential of generative AI: The next productivity frontier [Webinar]. McKinsey & Company. Link

65. Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv*. Link

66. Yee, L., & Chui, M. (2023). The economic potential of generative AI: The next productivity frontier [Webinar]. McKinsey & Company. Link

67. McKendrick, J. (2024). The C-Suite's hottest new job: The Chief AI Officer. *Forbes*. Link

68. World Economic Forum. (2023). Jobs of tomorrow: *Large language models and jobs*. Link

69. Demirci, O., Hannane, J., & Zhu, X. (2024). Research: How Gen AI is already impacting the labor market. *Harvard Business Review*. Link

70. Data & Society. (2024). Generative AI's labor impacts: *Part one—Hierarchy* [Event]. Link

71. Perrigo, B. (2022). Inside Facebook's African sweatshop. *TIME*. Link

72. Kumar, A., & Davenport, T. (2023). *How to make generative AI greener*. Harvard Business Review. Link

73. Milman, O. (2021). The hidden cost of silicon chips: Climate change's dirty secret. *The Guardian*. Link

74. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623*. Link

75. Kaufman, L., & de Chateauvieux, B. (2024). *How climate tech startups use generative AI to address the climate crisis*. AWS Startups Blog. Link

76. Crawford, K. (2024). Generative AI's environmental costs are soaring — and mostly secret. *Nature*. Link

77. von Eschenbach, W. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology*, 34: 1607–1622. Link

78. Levendowski, A. (2018). How copyright law can fix artificial intelligence's implicit bias problem. *Washington Law Review*, 93(2), 579–630. Link

79. Grynbaum, M. M., & Mac, R. (2023). The Times sues OpenAI and Microsoft over A.I. use of copyrighted work. *The New York Times*. Link

80. Reuel et al. (2024). Responsible AI in the Global Context: Maturity Model and Survey. Link

81. Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2021). Operationalising AI ethics: Barriers, enablers and next steps. *AI & Society*, 38, 411–423. Link

82. Smith, G., Luka, N., Osborne, M., Lattimore, B., Newman, J., Nonnecke, B. & Mittelstadt, B. (2025). Responsible Generative AI Use by Product Managers: Recoupling Ethical Principles and Practices. Link

83. Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2021). Operationalising AI ethics: Barriers, enablers and next steps. *AI & Society*, 38, 411–423. Link

84. Zhou, Y., & Chen, Y. (2023). AI ethics: From principles to practice. *AI & Society*, 38, 2693–2703. Link

85. Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J.-F., Brañas-Garza, P., Butera, L., Douglas, K. M., Everett, J. A. C., Gigerenzer, G., Greenhow, C., Hashimoto, D. A., Holt-Lunstad, J., Jetten, J., Johnson, S., Kunz, W. H., Longoni, C., ... Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*, 3(6). Link

86. Alba, D., & Yin, L. (2023). How generative AI can amplify racial, gender stereotypes. Bloomberg. Link

87. World Economic Forum. (2024). *Empowering AI leadership: An oversight toolkit for boards and business leaders*. Link

88. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. Link

89. Davenport, T., & Mittal, N. (2023, June 20). 13 principles for using AI responsibly. *Harvard Business Review*. Link

90. McKinsey & Company. (n.d.). *Responsible AI (RAI) principles*. Link

91. Meraw, S., & Wirtschafter, V. (2023). Putting research into practice: Brookings' approach to the responsible use of generative AI. Brookings Institution. Link

92. Smith, G., Luka, N., Osborne, M., Lattimore, B., Newman, J., Nonnecke, B. & Mittelstadt, B. (2025). Responsible Generative AI Use by Product Managers: Recoupling Ethical Principles and Practices. Link

93. Klyman, K. (2024). Acceptable use policies for foundation models. *arXiv*. Link

94. Goldman, P., & Baxter, K. (2023, February 7). Generative AI: 5 guidelines for responsible development. *Salesforce News*. Link

95. Salesforce. (n.d.). *Meet the Office of Ethical and Humane Use of Technology*. In *Ethics by Design* module. Retrieved November 25, 2024. Link

96. Goldman, P., & Niles, S. (2024). Salesforce joins national effort to build safe and trusted AI. *Salesforce News*. Link

97. This profile builds on the NIST AI Risk Management Framework, which defines seven characteristics of trustworthiness for artificial intelligence: valid and reliable; safe, secure and resilient; accountable and transparent; explainable and interpretable; privacy-enhanced; and fair with harmful biases managed. The framework breaks down the AI risk management process into four core functions: "govern," "map," "measure," and "manage," and each function is broken down further into categories and subcategories. The characteristics of trustworthiness are woven into these subcategories, particularly in “Govern 1.2” and “Measure 2”. While voluntary, they provide valuable shared language and sets of practices for organizations.

98. Principles include, for example: centering worker empowerment as their 'north star', as well as protecting labor and employment rights, supporting workers impacted by AI, and ensuring responsible use of worker data. The good practices include guidance such as providing appropriate training about AI systems in use to as broad a range of employees as possible throughout the enterprise, and not relying solely on AI and automated systems, or information collected through electronic monitoring, to make significant employment decisions.

99. Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible AI meets reality: *Practitioner perspectives on enablers for shifting organizational practices. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1)*, Article 7. Link

100. Ali, S. J., Christin, A., Smart, A., & Katila, R. (2023). Walking the walk of AI ethics: Organizational challenges and the individualization of risk among ethics entrepreneurs. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 10 pages. Link

101. Smith, G., Luka, N., Osborne, M., Lattimore, B., Newman, J., Nonnecke, B. & Mittelstadt, B. (2025). Responsible Generative AI Use by Product Managers: Recoupling Ethical Principles and Practices. Link

102. Microsoft. (2024). *2024 Responsible AI transparency report*. Link

103. Hazan, E., Madgavkar, A., Chui, M., Smit, S., Maor, D., Dandona, G. S., & Huyghues-Despointes, R. (2024). A *new future of work: The race to deploy AI and raise skills in Europe and beyond.* McKinsey Global Institute. Link

104. Smith, G., Luka, N., Osborne, M., Lattimore, B., Newman, J., Nonnecke, B. & Mittelstadt, B. (2025). Responsible Generative AI Use by Product Managers: Recoupling Ethical Principles and Practices. Link

105. Croak, M., & Gennai, J. (2022). An update on our work in responsible innovation. *The Keyword*. Link

106. Croak, M., & Gennai, J. (2023, June 22). Responsible generative AI: Google’s 3 emerging best practices. *The Keyword*. Link

107. Friedler, S., Singh, R., Blili-Hamelin, B., Metcalf, J., & Chen, B. J. (2023). *Recommendations for using red teaming for AI accountability.*

Berkeley
Haas

RE-AI.BERKELEY.EDU