2024 Tech Provider Top Trends: AI Safety

Published 12 December 2023 - ID G00804240 - 20 min read

By Analyst(s): Alan Antin, Tarun Rohilla

Initiatives: Business of High Tech; Digital Future; Technology Markets and Companies Insights for Investors

> The adoption of GenAI promises commercial gains, but it is also raising alarms about the harm it could cause. Product leaders need to invest in transparency and security to address the regulatory and market forces that are driving constant attention toward AI safety.

Overview

Opportunities

- As AI products proliferate in enterprise solutions, responsible and safe AI will increasingly be linked with transparency for both inputs and outputs as well as the need for trust and security. Explainability and transparency are crucial elements to facilitate adoption and provide safety guardrails.
- Providing data security measures for model creation can help safeguard AI products and provide clear audit trails for customers, regulators and other stakeholders. Inclusive in model security is developing clear guidelines for how training data is sourced for balance and rights usage.
- Providers can differentiate on the basis of enhanced security of proprietary information. Opportunities include launching new services or security products to safeguard customers, and strengthening value propositions by addressing the AI safety requirements that are surfacing. To ensure protection of proprietary information for customers, they can build in data-related security and other risk mitigation provisions into both enterprise and infrastructure applications with embedded AI.
- Providers can influence policymaking by collaborating with relevant organizations and preordain the areas that are critically relevant to technology and service providers'(TSPs') business and operating models.

Recommendations

Product leaders responsible for AI models and applications should:

- Build solutions that incorporate safety principles like responsible AI for model developers. Advance security objectives by ensuring input training data have clear usage rights, a documented model audit trail, and adherence to requirements for security and identity management.
- Invest in self-governance activities by belonging to industry consortiums, adhering to AI safety pledges, and implementing red teaming on own models. Speak openly and often about your participation in safety-related industry groups and your efforts to assure AI safety.
- Protect proprietary organizational and customer assets by training and upskilling your workforce and by acquiring solutions with relevant safety guardrails.
- Preempt regulatory and compliance issues by deploying AI trust, risk and security management (TRiSM) principles to manage technical and business requirements. Staying outfront is critical because regulation varies by jurisdiction.

What You Need to Know

This research takes a focused look into AI safety, which is one of the 10 highlighted trends in 2024 Tech Provider Top Trends.

AI has the potential to advance businesses and to create value like no other technology. AI can aid in the medical field with drug discovery as well as increase productivity and drive innovation. AI services alone are forecast to be a $443 billion market by 2027, with a compound annual growth rate (CAGR) of 16.9% from 2023 through 2027 (see Forecast Analysis: Artificial Intelligence Services, 2023-2027, Worldwide). And AI software spending is projected to grow to $298 billion by 2027, with a CAGR of 19.1% (see Forecast Analysis: Artificial Intelligence Software, 2023-2027, Worldwide).

Meanwhile, the mainstream arrival of generative AI (GenAI) has brought about new challenges. Consequently, some public and media interest in AI safety has been spurred by concerns about existential risks (e.g., “Could AI destroy the world?”), as well as concerns over societal disruptions arising from job losses and augmentations. These risks and disruptions need to be considered by governments and businesses.

For product leaders deploying AI, safety issues are pragmatic ones that must be addressed. These issues impact product roadmaps, the speed of adoption of AI use cases across business functions, enterprise competitiveness, and the ability to capture available market opportunities. At the same time, the widespread adoption of AI is spurring society, the tech community and governments to demand that this technology is deployed for the benefit of humankind, while maintaining a sharp focus on mitigating potential risks.

AI safety considerations are driving decision making for product leaders and will continue to do so in the years ahead. In particular, they should consider the AI safety issues impacting:

- **Builders:** This group is made up of tech providers that are building AI models to drive their applications and then reselling them directly to customers. AI builders put these models into production in commercial environments or make the systems available to their end customers who in turn build applications on top of them.
- **Users:** AI customers are an extremely broad constituency. Nearly all companies in developed countries are end users of technology that is infused with AI and often build their own models using applications designed by providers. As such, this is a technology built with democratization in mind, one that fulfills the needs for a wide spectrum of business functions (e.g., marketing, sales, supply chain, finance, non-IT and IT).
- **Regulators:** Government bodies address questions of public safety and assess the costs/benefits and ultimately the impact of AI on society. Ethical factors weigh into these considerations, reflecting the values of a society. This in turn creates a further challenge for product leaders because of the fragmented nature of rules originating in different regions.

AI safety is trending as governments begin to understand the potential of AI and the harm it can do to society. AI builders and researchers, meanwhile, are appealing for regulation, and government and business leaders are asking ethical questions about the technology. In light of these forces, enterprises must prepare for evolving demands around responsible and safe AI. Such demands are coming from those seeking safeguards with this advancing technology: governments, nongovernmental organizations (NGOs) and industry consortiums as well as several tech companies that are attempting to lead by example.

Profile: AI Safety

Description:

AI safety comprises a series of steps that providers of AI-based tools and capabilities take to address AI-related concerns while continuing to spur adoption and innovation. These steps address issues of:

- Tracking of individual behaviors
- Biases in models
- Models producing output that states false information but appears to come from legitimate sources (hallucinations)
- Eroding trust in information through deepfakes (text, audio, image, video or a combination)
- Global security threats by state and nonstate actors

Additional issues include needing to protect company proprietary data, wanting to leverage best-in-class training data while having clear ownership rights on that data, ensuring that the AI is not harming a brand's reputation, and implementing a robust set of security measures for models and their supporting compute infrastructure.

The steps that tech providers are taking to mitigate these issues include:

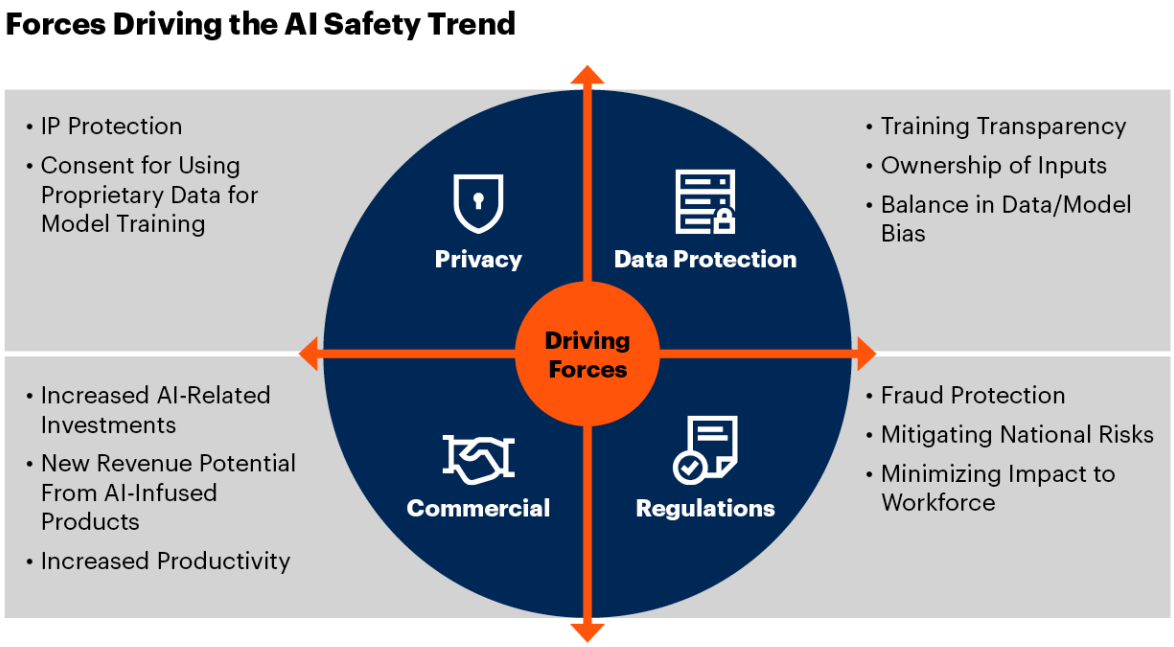
- Increasing transparency methods for model outputs and input
- Developing warranties and guarantees that usage of their products is protected
- Implementing security measures for those AI engineers developing models (e.g., Microsoft Azure AI content safety offerings and other explainable AI products)
- Training staff to be overall stewards of company IP whether building or using AI

The underlying force at work is responsible AI, which is an umbrella term for many aspects of making the right business and ethical choices when an organization adopts AI (for more insights on responsible AI, see Emerging Tech Impact Radar: Artificial Intelligence).

Why Trending:

Responsible and safe AI is a trend that impacts many roles. It is especially a key concern for product leaders as, according to our product management survey, AI/GenAI is the most impactful trend affecting product management teams over the next 12 months. AI safety is meaningful right now for product leaders due to four driving forces impacting those product leaders responsible for AI in their organizations (see Figure 1).

Figure 1: Forces Driving the AI Safety Trend



1. **Privacy rights.** AI is incredibly good at identifying and predicting patterns around individual behaviors, or personal image identification through facial recognition systems. [1] While convenience and personalization may be enhanced through such uses, there is also a significant concern that using personal data in these ways can also be easily manipulated for nefarious purposes. User agreements must be reflective of how end-user data will be used and explicit permissions obtained. [2]

2. **Protection of data and intellectual property (IP).** Certain AI models have been trained without explicit agreement from original data sources using techniques such as web scraping. Companies are taking notice and developing strategies to protect their IP through negotiated agreements for data and/or they are seeking legal recourse. Additionally, enterprises are looking to build specific models with proprietary training data that cannot be leaked into public domains.

3. **Government regulations.** Government leaders around the world have taken notice of the transformative nature of AI and want to put safeguards in place while not hampering innovation or potential societal benefits. Unlike other technologies that function within the set of dimensions as defined in their source code, AI outputs may not provide a predefined and expected outcome. The autonomy that AI possesses and the black-box nature of the models and algorithms make this a unique challenge for regulators (see Top Trends in AI Public Policy and Regulations for 2024).

4. **Commercial imperatives**. Enterprises are adopting AI-infused products as this transformative technology touches multiple functions inside businesses, creating new revenue opportunities and productivity gains (see Forecast Analysis: Artificial Intelligence Software, 2023-2027, Worldwide). According to the 2023 Gartner CIO and Technology Executive Survey, before ChatGPT made headlines, 32% of CIOs and technology executives said their enterprise would spend the largest amount of new or additional funding in 2023 on AI/machine learning in general (e.g., not just GenAI). Another indicator of the commercial interest in AI products is the heightened activity in venture capital (VC) markets. VC investment has increased 425% in GenAI since 2020. [3]

The hyperattention resulting from recent advances in AI solutions leveraging large language models (LLMs) is accelerating the AI safety trend due to concerns about deepfakes, the ease with which misinformation can be spread and the potential to harm company IP. As per the 2023 Gartner IT Executives Webinar Poll, the most worrisome risks of GenAI include data privacy (42%), hallucinations (14%), misuse (12%), security (13%), bias and unfairness (10%), and black-box responses (7%). While some AI products are mature and have been in production for years, other emerging GenAI use cases are raising a host of challenges and risks as organizations seek to harvest the benefits.

Key AI safety topics that product leaders need to be aware of now and in the years to come include the following:

- When is it important for end users to understand that they are interacting with robots or AI? Deepfakes in public elections is an extreme on one end, but also the idea of AI transparency has ramifications in business use cases pertaining to user experiences, content production, simulation and search marketing (for more use cases, see Emerging Tech: Top 4 Security Risks of GenAI and Emerging Tech: Primary Impact of Generative AI on Business Use Cases).

- What measures do tech providers and their customers need to take to enjoy the benefits of AI, while also protecting company and customer information from leaking out as training data?
- How do tech provider employees need to be trained to maximize the benefits of AI and mitigate the risks of inadvertent exposures of proprietary company information?
- When tech providers are building AI models, either those that will be publicly available as open source, domain-specific or privately operated, should developers have the rights to use available training data without the consent of the owners?
- What measures can be implemented to ensure equality and algorithmic accountability? When developing AI models, the training data may intrinsically introduce biases in the outputs. Recent history with social media platforms, credit loan AI systems and other use cases have shown the downside of the technology.
- How can the outputs from AI be understood by users? In the case of LLMs, AI builders will be stymied in their attempts to explain how neural networks arrive at answers. As the model power scales from gains in compute, parameter size, and model efficiencies, it will be difficult to impossible for explainability to keep up.
- How should AI be used safely within tech providers? Providers need to make sure that code generated or assisted by AI does not introduce new forms of vulnerabilities.

Implications:

AI Safety Is Fueling Government Regulations

Many countries have commenced their work on AI-specific regulations. For instance, the EU member countries are commenting on a draft AI Act and will likely adopt this legislation as law by the end of 2023. The AI Act is a risk-based system that classifies AI systems by four levels: prohibited, regulated high risk, transparent limited risk, and low and minimal risk. Prohibited systems are those that manipulate human behavior or exploit vulnerabilities. High-risk AI systems are those used for biometric identification, recruitment, credit scoring, education or healthcare. The high-risk AI systems must comply with strict rules on data quality, transparency, accuracy, robustness and security. They must also undergo a conformity assessment before being placed on the market.

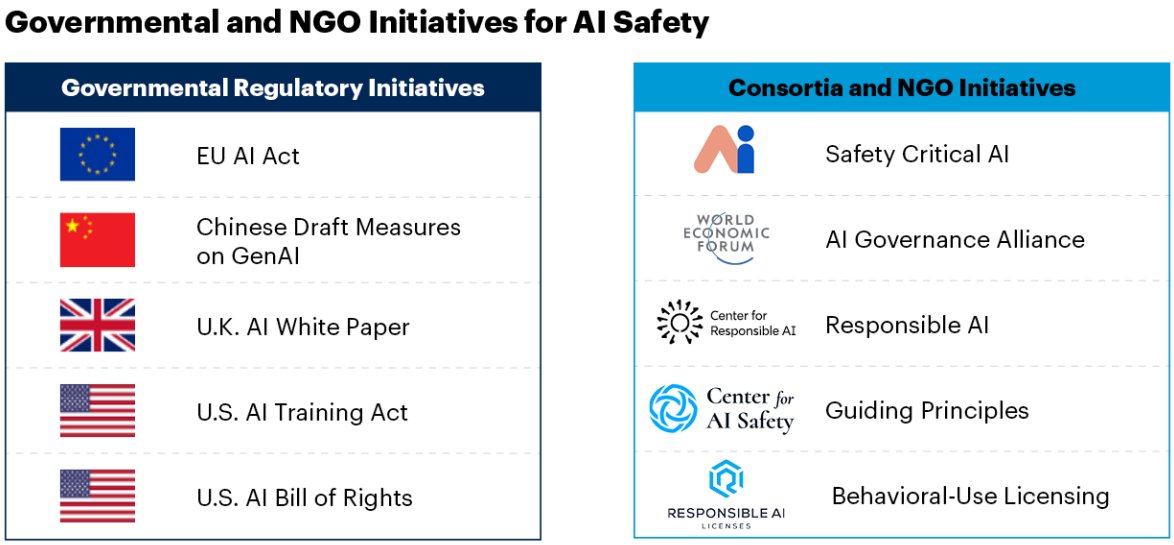
The EU's AI Act also aims to establish a governance structure, a European Artificial Intelligence Board (EAIB), for the enforcement of its rules. Sanctions and remedies for noncompliance may include fines of up to 6% of annual worldwide turnover or €30 million (whichever is higher) for serious infringements. Similarly, the U.S. Congress is discussing an AI Bill of Rights and seeks to put it into law in short order, and President Biden has issued an Executive Order with guidelines. China's Interim Administrative Measures for Generative AI Services came into effect in August 2023. These advancements signify an increased focus of governmental organizations on the implications and risks of AI.

Prior technology safety initiatives for privacy resulted in wide-sweeping measures being implemented (e.g., the General Data Protection Regulation and California Consumer Privacy Act). But other initiatives seeking to regulate technology (e.g., in social media) have stalled. That's because regulators struggle with the balancing act of enabling free speech versus restricting potentially harmful algorithms, as well as unleashing innovation and competition versus considering the impacts of new technology on society. [4]

In the realm of AI safety, it is likely that regulations will come sooner rather than later. First, there is the fear of the "unknown unknowns" by lawmakers. This is coupled with the leading voices in AI engineering asking governments to regulate or pause developments in the industry. [5] Such regulations mean that certain model developers will require a license, and that safety approvals will be needed from sanctioning agencies charged with regulating these models. AI builders will need to account for these approvals in their development and go-to-market (GTM) timelines. For key AI infrastructure, export controls will restrict advanced computing technologies being sold to certain countries. [6]

While governments are preparing to enact rules that will spell out the bare minimum requirements for AI safety compliance, industry consortiums and NGOs are also promoting measures to ensure ethical and safe development (see Figure 2).

Figure 2: Governmental and NGO Initiatives for AI Safety



AI Safety Will Drive Transparency for Data and IP Protection

High-profile AI builders have been in the spotlight because their LLMs have been trained on data without the explicit permission of the original creators. To offset this concern, training data will increasingly be licensed for a fee or via partnership. AI builders must also acknowledge and document the training data and, in certain instances, partner to pay for the clear rights to these inputs. Users must investigate input sources to understand the outputs and the potential risks. To further transparency, AI builders will incorporate digital watermarks in content to distinguish human versus machine-generated outputs.

AI systems may also ingest company data and can expose that information publicly if security measures are not in place. Enterprises will need to implement AI TRiSM (see Market Guide for AI Trust, Risk and Security Management) and employee training programs to ensure that staff are safely working with systems. Further upskilling will be needed for any staff using AI products to ensure compliance and manage risks.

AI Safety Will Further Propel Privacy Rights

AI builders must recognize that when using personal information for training models, there must be clear opt-out and usage-related provisions. Systems that provide monitoring or recognition systems will need to ensure documented consent of participants. AI systems that use algorithms for decision making, or augmenting human decision making, will need widely available training inputs to represent the marketplace in a fair and balanced way. Tech providers of such systems will need to document inputs and training data to answer calls for algorithmic accountability.

AI Safety Will Not Impede Commercial Imperatives

The underlying force that steers businesses to seek profit margin improvements and/or revenue growth will continue to drive companies to capture the benefits of deploying AI. Automation and productivity gains are commercial imperatives that accelerate the adoption of AI products. New revenue streams from AI products and services are also a commercial imperative that drives AI offerings.

Each of these imperatives is intrinsically linked with responsible and safe AI. When asked if the benefits of GenAI outweigh its risks, 68% of executive leaders said they see more benefits than risks (see Executive Pulse: AI Investment Gets a Boost From ChatGPT Hype). Through the adoption of various AI use cases, businesses will seek to reap commercial gains and competitive advantage. In the race to realize these strategic gains and to be innovative and competitive, tech providers must also be mindful of the potential risks. The market leaders will integrate safety measures proactively into products to participate as fully as possible in the benefits.

Impact on Technology and Service Provider Business Operations:

One of the defining attributes of a top trend is its multifaceted impact on different parts of a TSP's business operations. That is also the case for AI safety, whose impact across the analyzed six domains is summarized in Table 1.

Table 1: Impact of AI Safety on TSP Business Operations

(Enlarged table in Appendix)

Business Operations	Scale of Impact	Scope of Impact
Products and Services	High	▪ AI is infused in applications for functions across the enterprise, and so the concerns about building in the needed facets for safety and compliance will increasingly impact tech providers. ▪ Product and service provider messaging and GTM strategy will require adjustments to reassure potential buyers of adherence to AI safety principles.
Customers and Buyers	Very High	▪ Customers are becoming more aware of the various risks of AI and will seek reassurances such as indemnification provisions, model transparency, and guarantees of validated usage rights. ▪ Providers will craft value propositions to cater to buyers' behaviors and expectations toward responsible and safe AI. ▪ Buyers will assess risk versus reward for potential reputation and brand damages if the AI goes awry.
Operations and Processes	Very High	▪ Building in varied levels of explainability and transparency will assuage some AI concerns while also helping to accelerate adoption. ▪ Providers will be required to update their operating model to account for increased scrutiny and to mobilize necessary resources.
Competitive Landscape	High	▪ Larger providers and frontier model developers will have a competitive advantage in distribution and resources for AI safety provisions. ▪ New entrants in application and domain-specific products, which have addressed safety, will gain an edge over generic solutions. ▪ Niche providers will emerge in the AI safety domain, launching AI-safety-focused products and services.
Partners and Ecosystems	Medium	▪ Product leaders will forge new partnerships with industry consortiums and NGOs to remain at parity with industry norms. ▪ TSPs that are seeking to build leading-edge AI safety measures in products will need partnerships with specialized experts in transparency, explainability and related responsible and ethical AI development methodologies.
Talent and Resources	Very High	▪ Training employees on AI safety will be fluid as rules will vary by region, policies will vary by company, and knowledge base will vary by workers. ▪ It will be difficult to source and hire those data and security experts who can ensure compliance.

Definitions for scale of impact:
Very high = above 75% of business operations impacted for all tech providers
High = 50% to 75% of business operations impacted for all tech providers
Medium = 30% to 50% of business operations impacted for all tech providers
Low = below 30% of business operations impacted for all tech providers

Source: Gartner (December 2023)

Actions:

- Monitor the regulatory environment in major geographic regions: the EU, the U.S. and China. Ensure that compliance and legal teams are tracking AI rules as they emerge and evolve. Penalties for noncompliance will be severe, and AI providers need to remain current on rules in the markets where they operate.
- Support and/or track self-governance activities such as the Frontier Model Forum, the AI Governance Alliance, Responsible AI Institute and the AI Now Institute.
- Develop a strategy for relevant input and output transparency. In GenAI, output transparency may mean digital watermarks for content deliverables. Early examples of this include SynthID from Google Cloud and DeepMind.

- Invest in explainability, transparency and interpretability functionality to provide reassurance related to safety concerns and to speed end-user adoption. Explore reverse-engineering what your neural network model is actually doing, and further your ability to explain outputs.
- Assess AI value propositions for safety with a lens on model transparency, traceability, interpretability and explainability aspects.
- Pressure-test models and applications with vigorous red teaming prior to general availability. Follow best-practice security processes in AI TRiSM (see Top Strategic Technology Trends for 2024: AI Trust, Risk and Security Management).
- Develop guiding principles for ethical AI that are actionable and codable in practice. Examples of this include Google's AI Principles [7] and Anthropic's Constitutional AI. [8]
- Provide guarantees of data protection to reassure customers that model inputs are safe. Early examples include the Microsoft Copilot Copyright Commitment and Adobe Firefly's copyright infringement protection. [9]

Evidence

2023 Gartner Product Management Branded Survey. This survey was conducted to understand the needs, challenges and key initiatives of people in product management roles from June through August 2023. In total, 283 respondents were interviewed across the U.S. (n = 83), Canada (n = 35), the U.K. (n = 31), Germany (n = 28), China (n = 27), France (n = 25), India (n = 23), Australia (n = 18) and Hong Kong (n = 14). Small base sizes of fewer than 30 respondents should be interpreted with caution. To enable key trends to be compared and contrasted, we established quotas on key organizational and respondent characteristics. Qualifying organizations operate in high-tech industries (cloud services; software; communications equipment; carriers; devices and computer infrastructure; technology and business services; and semiconductors) with anticipated enterprisewide annual revenue for 2023 of more than $50 million or equivalent. Qualified participants have the title of director or equivalent and above with product development/engineering or product management as their primary job function/department. Disclaimer: The results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

2023 Gartner CIO and Technology Executive Survey. This survey was conducted to help CIOs and technology executives overcome digital execution gaps by empowering and enabling an ecosystem of internal and external digital technology producers. It was conducted online from 2 May through 25 June 2022 among Gartner Executive Programs members and other CIOs. Qualified respondents are each the most senior IT leader (e.g., CIO) for their overall organization or some part of their organization (for example, a business unit or region). The total sample is 2,203 respondents, with representation from all geographies and industry sectors (public and private). Disclaimer: Results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

2023 Gartner IT Executives Webinar Poll. This webinar poll was conducted to help IT leaders using or building GenAI apps to understand market dynamics and evaluate emerging GenAI TRiSM technology and providers that address new risks. It was conducted during the webinar hosted in August 2023. Qualified respondents were those who signed up and attended the webinar hosted by Gartner. The total sample was 713 respondents who answered the poll question: Which risks of GenAI are you most worried about? Options provided as responses were: bias and unfairness, black-box responses, data privacy, hallucinations, misuse, security, and others. Disclaimer: Results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents.

[1] 7 Biggest Privacy Concerns Around Facial Recognition Technology, Civil Liberties Union for Europe.

[2] Facial Recognition Technology and Privacy Concerns, ISACA.org.

[3] Generative AI: The New Frontier For VC Investment, Forbes.

[4] Don’t Kill Innovation With Excessive Regulation, Scientific American.

[5] Pressured by Biden, A.I. Companies Agree to Guardrails on New Tools, The New York Times.

[6] U.S. Curbs AI Chip Exports From Nvidia and AMD to Some Middle East Countries, Reuters.

[7] Google AI Responsibility, Google.

[8] Constitutional AI: Harmlessness From AI Feedback, Anthropic.

[9] Adobe Is So Confident Its Firefly Generative AI Won't Breach Copyright It'll Cover Your Legal Bills, Fast Company.

Recommended by the Authors

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

Applying AI — Governance and Risk Management

A Comprehensive Guide to Responsible AI

What Executives Need to Do to Support the Responsible Use of AI

Top 5 Priorities for Managing AI Risk Within Gartner's MOST Framework

Market Guide for AI Trust, Risk and Security Management

Quick Answer: Will AI Create an Existential Crisis for Humanity?

Table 1: Impact of AI Safety on TSP Business Operations

Business Operations	Scale of Impact	Scope of Impact
Products and Services	High	■ AI is infused in applications for functions across the enterprise, and so the concerns about building in the needed facets for safety and compliance will increasingly impact tech providers. ■ Product and service provider messaging and GTM strategy will require adjustments to reassure potential buyers of adherence to AI safety principles.
Customers and Buyers	Very High	■ Customers are becoming more aware of the various risks of AI and will seek reassurances such as indemnification provisions, model transparency, and guarantees of validated usage rights. ■ Providers will craft value propositions to cater to buyers' behaviors and expectations toward responsible and safe AI. ■ Buyers will assess risk versus reward for potential reputation and brand damages if the AI goes awry.

Operations and Processes	Very High	■ Building in varied levels of explainability and transparency will assuage some AI concerns while also helping to accelerate adoption. ■ Providers will be required to update their operating model to account for increased scrutiny and to mobilize necessary resources.
Competitive Landscape	High	■ Larger providers and frontier model developers will have a competitive advantage in distribution and resources for AI safety provisions. ■ New entrants in application and domain-specific products, which have addressed safety, will gain an edge over generic solutions. ■ Niche providers will emerge in the AI safety domain, launching AI-safety-focused products and services.
Partners and Ecosystems	Medium	■ Product leaders will forge new partnerships with industry consortiums and NGOs to remain at parity with industry norms.

		■ TSPs that are seeking to build leading-edge AI safety measures in products will need partnerships with specialized experts in transparency, explainability and related responsible and ethical AI development methodologies.
Talent and Resources	Very High	■ Training employees on AI safety will be fluid as rules will vary by region, policies will vary by company, and knowledge base will vary by workers. ■ It will be difficult to source and hire those data and security experts who can ensure compliance.

Definitions for scale of impact:
Very high = above 75% of business operations impacted for all tech providers
High = 50% to 75% of business operations impacted for all tech providers
Medium = 30% to 50% of business operations impacted for all tech providers
Low = below 30% of business operations impacted for all tech providers

Source: Gartner (December 2023)