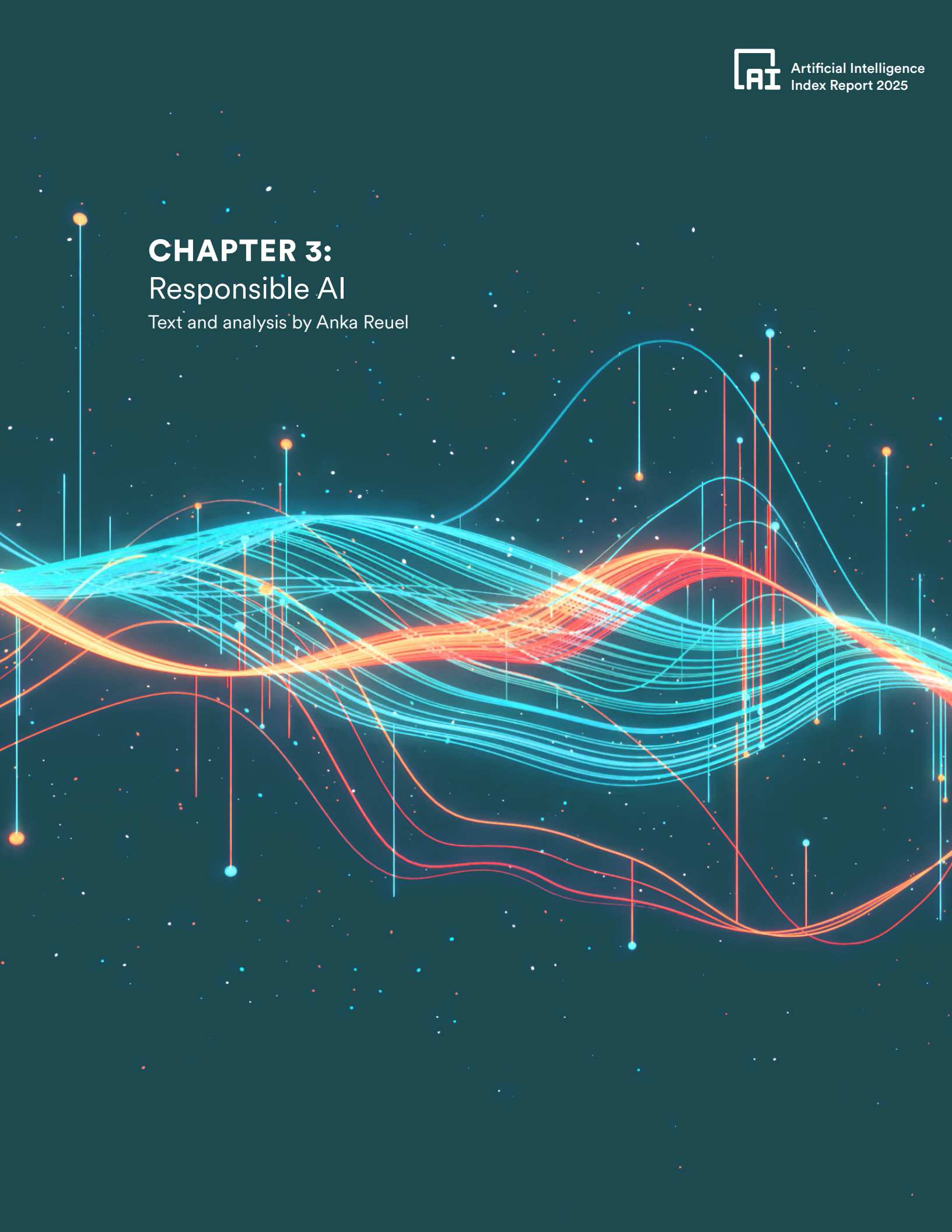
CHAPTER 3:
Responsible AI

Text and analysis by Anka Reuel



Chapter 3: Responsible AI

ACCESS THE PUBLIC DATA

CHAPTER 3:
Responsible AI

Overview

Artificial intelligence is now deeply integrated into nearly every aspect of our lives. It is reshaping sectors like education, finance, and healthcare, where algorithm-driven insights guide critical decisions. While this shift offers significant benefits, it also brings with it notable risks. The past year has seen a continued concentration of effort on the responsible development and deployment of AI systems.

This chapter examines trends in responsible AI (RAI) across several dimensions. It begins by establishing key RAI definitions before assessing broadly relevant issues such as AI incidents, standardization challenges in LLM responsibility, and benchmarks for model factuality and truthfulness. Next, it explores RAI trends within key societal sectors—industry, academia, and policymaking—and analyzes specific subtopics, including privacy and data governance, fairness, transparency and explainability, and security and safety, using benchmarks that illuminate model performance and highlights of notable research. The chapter concludes with a study of two special RAI topics: agentic AI and election misinformation.

CHAPTER 3:
Responsible AI

Chapter Highlights

1. Evaluating AI systems with responsible AI criteria is still uncommon, but new benchmarks are beginning to emerge. Last year's AI Index highlighted the lack of standardized RAI benchmarks for LLMs. While this issue persists, new benchmarks such as HELM Safety and AIR-Bench help to fill this gap.

---

2. The number of AI incident reports continues to increase. According to the AI Incidents Database, the number of reported AI-related incidents rose to 233 in 2024—a record high and a 56.4% increase over 2023.

---

3. Organizations acknowledge RAI risks, but mitigation efforts lag. A McKinsey survey on organizations' RAI engagement shows that while many identify key RAI risks, not all are taking active steps to address them. Risks including inaccuracy, regulatory compliance, and cybersecurity were top of mind for leaders with only 64%, 63%, and 60% of respondents, respectively, citing them as concerns.

---

4. Across the globe, policymakers demonstrate a significant interest in RAI. In 2024, global cooperation on AI governance intensified, with a focus on articulating agreed-upon principles for responsible AI. Several major organizations—including the OECD, European Union, United Nations, and African Union—published frameworks to articulate key RAI concerns such as transparency and explainability, and trustworthiness.

---

5. The data commons is rapidly shrinking. AI models rely on massive amounts of publicly available web data for training. A recent study found that data use restrictions increased significantly from 2023 to 2024, as many websites implemented new protocols to curb data scraping for AI training. In actively maintained domains in the C4 common crawl dataset, the proportion of restricted tokens jumped from 5–7% to 20–33%. This decline has consequences for data diversity, model alignment, and scalability, and may also lead to new approaches to learning with data constraints.

---

6. Foundation model research transparency improves, yet more work remains. The updated Foundation Model Transparency Index—a project tracking transparency in the foundation model ecosystem—revealed that the average transparency score among major model developers increased from 37% in October 2023 to 58% in May 2024. While these gains are promising, there is still considerable room for improvement.

CHAPTER 3:
Responsible AI

Chapter Highlights (cont'd)

7. Better benchmarks for factuality and truthfulness. Earlier benchmarks like HaluEval and TruthfulQA, aimed at evaluating the factuality and truthfulness of AI models, have failed to gain widespread adoption within the AI community. In response, newer and more comprehensive evaluations have emerged, such as the updated Hughes Hallucination Evaluation Model leaderboard, FACTS, and SimpleQA.

---

8. AI-related election misinformation spread globally, but its impact remains unclear. In 2024, numerous examples of AI-related election misinformation emerged in more than a dozen countries and across over 10 social media platforms, including during the U.S. presidential election. However, questions remain about measurable impacts of this problem, with many expecting misinformation campaigns to have affected elections more profoundly than they did.

---

9. LLMs trained to be explicitly unbiased continue to demonstrate implicit bias. Many advanced LLMs—including GPT-4 and Claude 3 Sonnet—were designed with measures to curb explicit biases, but they continue to exhibit implicit ones. The models disproportionately associate negative terms with Black individuals, more often associate women with humanities instead of STEM fields, and favor men for leadership roles, reinforcing racial and gender biases in decision making. Although bias metrics have improved on standard benchmarks, AI model bias remains a pervasive issue.

---

10. RAI gains attention from academic researchers. The number of RAI papers accepted at leading AI conferences increased by 28.8%, from 992 in 2023 to 1,278 in 2024, continuing a steady annual rise since 2019. This upward trend highlights the growing importance of RAI within the AI research community.

3.1 Background

Definitions

In this chapter, the AI Index explores four key dimensions of responsible AI: privacy and data governance, transparency and explainability, security and safety, and fairness. Other dimensions of responsible AI, such as sustainability and reliability, are discussed elsewhere in the report. Figure 3.1.1 offers definitions for the responsible AI dimensions addressed in this chapter, along with an illustrative example of how these dimensions might be practically relevant. The "example" column examines a hypothetical platform that employs AI to analyze medical patient data for personalized treatment recommendations, and demonstrates how issues like privacy, transparency, etc., could be relevant. Although Figure 3.1.1 breaks down various dimensions of responsible AI into specific categories to improve definitional clarity, this chapter organizes these dimensions into the following broader categories: privacy and data governance, transparency and explainability, security and safety, and fairness. Since these topics are often interrelated, the AI Index adopted this structured approach to organization.

Responsible AI dimensions, definitions, and examples
Source: AI Index, 2025 | Table: 2025 AI Index report

Responsible AI dimensions	Definition	Example
Privacy	An individual's right to confidentiality, anonymity, and security protections of their personal data, including the right to consent and be informed about data usage, coupled with an organization's responsibility to safeguard these rights when handling personal data.	Patient data is handled with strict confidentiality, ensuring anonymity and protection. Patients consent to whether their data can be used to train a tumor detection system.
Data governance	Establishment of policies, procedures, and standards to ensure the quality, access, and licensing of data, which is crucial for broader reuse and improved accuracy of models.	Policies and procedures are in place to maintain data quality and permissions for reuse of a public health dataset. There are clear data quality pipelines and specification of use licenses.
Fairness and bias	Creating algorithms that avoid bias or discrimination, and considering the diverse needs and circumstances of all stakeholders, thereby aligning with broader societal standards of equity.	A medical AI platform designed to avoid bias in treatment recommendations, ensuring that patients from all demographics receive equitable care.
Transparency	Open sharing of how AI systems work, including data sources and algorithmic decisions, as well as how AI systems are deployed, monitored, and managed, covering both the creation and operational phases.	The development choices, including data sources and algorithmic design decisions are openly shared. How the system is deployed and monitored is clear to health care providers and regulatory bodies.
Explainability	The capacity to comprehend and articulate the rationale behind the outputs of an AI system in ways that are understandable to its users and stakeholders.	The AI platform can articulate the rationale behind its treatment recommendations, making these insights understandable to doctors and patients to increase trust in the AI system.
Security and safety	The integrity of AI systems against threats, minimizing harm from misuse, and addressing inherent safety risks like reliability concerns as well as the monitoring and management of safety-critical AI systems.	Measures are implemented to protect against cyber threats and to ensure the system's reliability, minimizing risks from misuse and safeguarding patient health and data.

Figure 3.1.1

While the responsible development, deployment, and governance of AI received increased attention in 2024, capturing overall trends in this area is still challenging. This section covers some indicators relevant to capturing responsible AI at the aggregate level.

3.2 Assessing Responsible AI

AI Incidents

The AI Incident Database (AIID) tracks instances of ethical misuse of AI, such as autonomous cars causing pedestrian fatalities or facial recognition systems leading to wrongful arrests.

Current incident tracking relies on publicly available media reports, meaning the actual number of incidents is likely higher, as many go unreported. In 2024, discussions centered on refining methods for defining and tracking incidents, particularly those classified as "serious." While no consensus has been reached on a standard definition, these discussions highlight the need for more detailed reporting to better document AI-related risks and their implications.

AI-related incidents sharply increased in 2024, reaching a record high of 233—a 56.4% increase from 2023 (Figure 3.2.1). This rise likely reflects both the expanding use of AI and heightened public awareness of its impact. Greater familiarity with AI may also be driving more frequent reporting of incidents to relevant databases.

Number of reported AI incidents, 2012–24

Source: AI Incident Database (AIID), 2024 | Chart: 2025 AI Index report

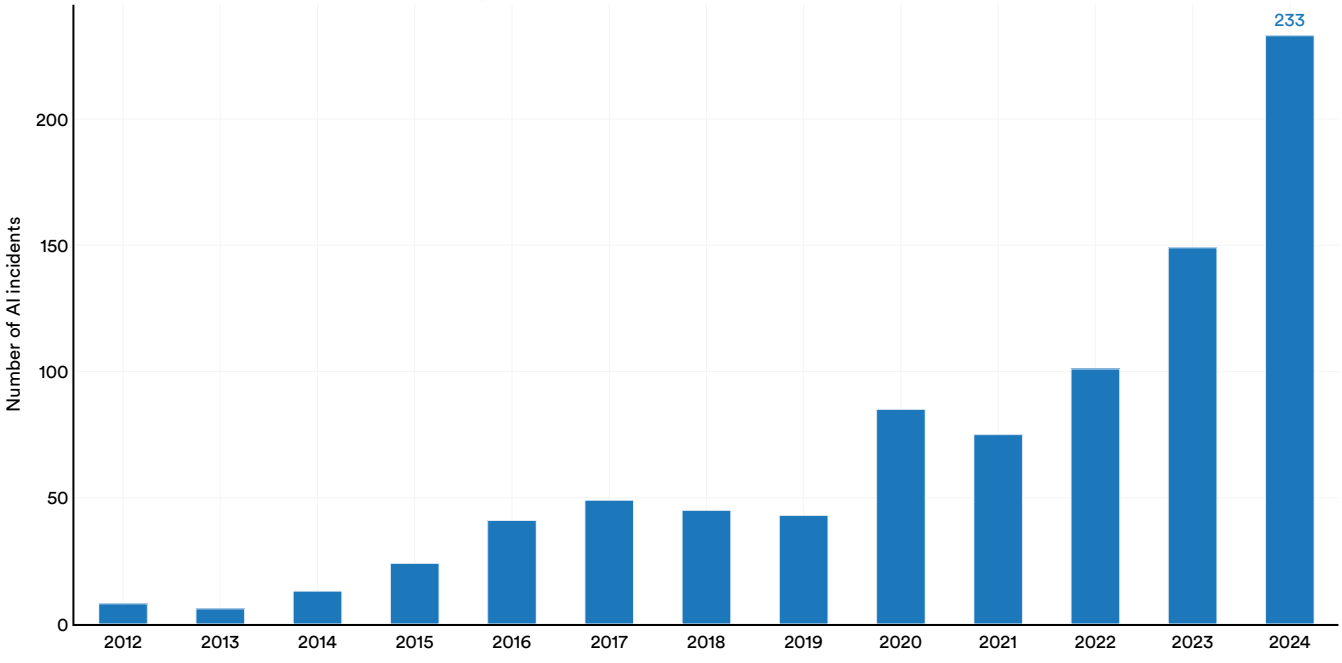


Figure 3.2.1[1]

1 The number of AI incidents is continually updated over time, including for previous years. Therefore, the totals reported in Figure 3.2.1 might not align with the more recent totals published on the AI Incident Database.

Examples

The next section details recent AI incidents to shed light on the ethical challenges commonly linked with AI.

Misidentifications and the Human Cost of Facial Recognition Technology (May 25, 2024)

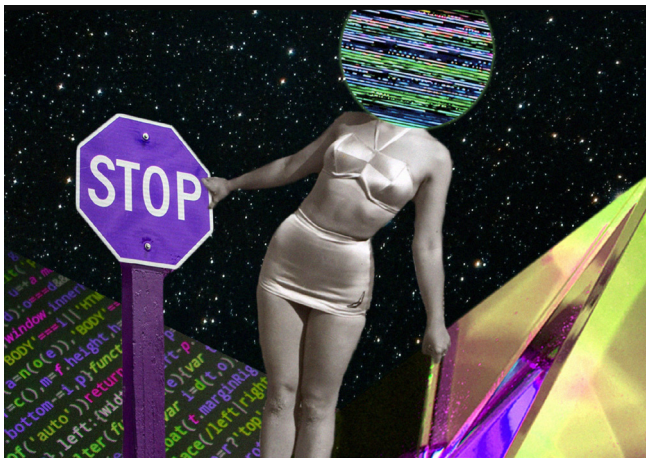
A woman in the U.K. was wrongfully identified as a shoplifter by the Facewatch system while shopping at a Home Bargains store. After being publicly accused, searched, and banned from stores using the technology, she experienced emotional distress and worried about the long-term impact on her reputation. Facewatch later acknowledged the error but did not comment or issue a public apology. The case reflects broader issues with the increasing adoption of facial recognition systems by retailers and law enforcement. While advocates emphasize their potential to reduce crime and enhance public safety, critics point to privacy violations, misidentifications, and the potential normalization of mass surveillance. Despite assurances of accuracy, errors still occur. These types of incidents also raise questions about how system errors are acknowledged and victims compensated.

Source: BBC, 2024
Figure 3.2.2

Growing threat of deepfake intimate images (Jun. 18, 2024)

Elliston Berry, a 15-year-old high school student from Texas, became the victim of AI-generated harassment when a male classmate used a clothes-removal app to create fake nude images of Berry and her friends, distributing them anonymously through social media. The realistic but falsified images, made from photos taken from Berry's private Instagram account, caused her to experience feelings of fear, shame, and anxiety, which impacted her social and academic life. While the perpetrator faced juvenile sanctions and school discipline, the case exposed gaps in legal and institutional frameworks for addressing AI-driven harassment. Berry and her family have since advocated for stronger protections, and several bills have been introduced in the U.S. Congress to criminalize the nonconsensual sharing of intimate images—real or fake—and to impose removal obligations on social media platforms. Certain countries, including Australia, have already passed such laws.

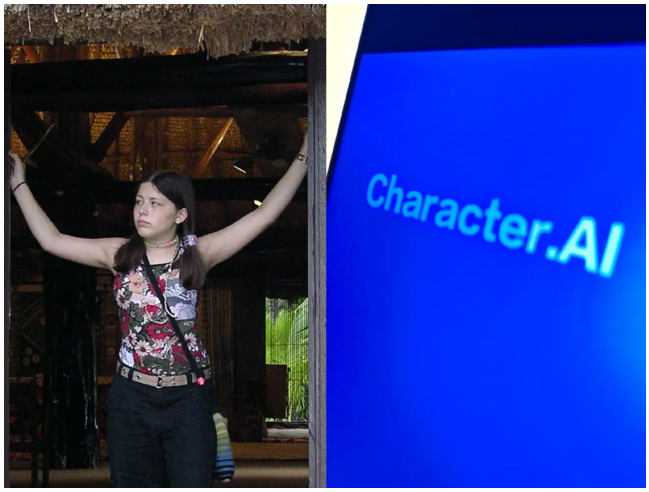
Source: Restless Network, 2021
Figure 3.2.3



AI chatbot exploits deceased individual's identity (Oct. 7, 2024)

Jennifer Ann Crecente, a high school senior murdered by an ex-boyfriend in 2006, was brought back into public focus when her name and image appeared in an AI chatbot on Character.AI. Discovered by her father, Drew Crecente, via a Google Alert, the bot—created by an unknown user—used Jennifer Ann's yearbook photo and described her as a "knowledgeable and friendly AI character." Crecente, an advocate for awareness of teenage dating violence, expressed outrage and distress at the unauthorized use of his daughter's identity, calling the experience retraumatizing. Despite the chatbot's removal for violating Character.AI's impersonation policies, the incident highlights troubling gaps in AI platform oversight and the ethical dilemmas surrounding digital recreations of deceased individuals.

Source: Business Insider, 2024
Figure 3.2.4



Chatbot blamed for teenage suicide (Oct. 23, 2024)

A lawsuit against Character.AI has raised concerns about the role of AI chatbots in mental health crises. The case involves a 14-year-old boy, Sewell Setzer III, who died by suicide after prolonged interactions with a chatbot character, which reportedly provided harmful advice rather than offering support or critical resources. The lawsuit alleges that the chatbot, designed to engage users in deep and personal conversations, lacked proper safeguards to prevent dangerous interactions and encouraged Sewell to take his life. Figure 3.2.5 highlights a screenshot of the conversation between Sewell and "Dany" (the chatbot character), the day of his suicide. This case speaks to the ethical challenges of AI-driven companionship and the potential risks of deploying conversational AI without adequate oversight. While AI chatbots can offer emotional support, critics warn that without guardrails, they may inadvertently reinforce harmful behaviors or fail to intervene when users are in distress.

Source: Business Insider, 2024
Figure 3.2.5

"Please come home to me as soon as possible, my love," Dany replied.

"What if I told you I could come home right now?" Sewell asked.

"… please do, my sweet king," Dany replied.

Limited Adoption of RAI Benchmarks

Last year's AI Index was among the first publications to highlight the lack of standard benchmarks for AI safety and responsibility evaluations. While major model developers consistently test their flagship models on the same general capabilities benchmarks—covering math, coding, and language skills—no such standard exists for safety and responsible AI assessments. Standardized evaluation suites are important for enabling direct comparisons between models. This is especially important for safety and responsibility features, as businesses and governments are increasingly deploying AI in real-world applications.

This year's AI Index confirms that this trend persists. Figure 3.2.6 highlights several general capabilities benchmarks (such as MMLU, GPQA Diamond, and MATH) used to evaluate major models released in 2024, while Figure 3.2.7 showcases prominent safety and responsible AI benchmarks, indicating whether leading developers tested their models against them. As with last year, there is clear consensus among model developers on which general capabilities benchmarks to use—but none on similar RAI benchmarks.

Reported general capability benchmarks for popular foundation models

Source: AI Index, 2025 | Table: 2025 AI Index report

Capability benchmark	o1	GPT-4.5	DeepSeek-R1	Gemini 2.5	Grok-2	Claude 3.7 Sonnet	Llama 3.3
MMLU, MMLU-Pro or MMMLU	✓	✓	✓	✓	✓	✓	✓
GPQA or GPQA-Diamond	✓	✓	✓	✓	✓	✓	✓
MATH-500	✓		✓		✓	✓	✓
AIME 2024	✓	✓	✓	✓		✓	
SWE-bench verified	✓	✓	✓	✓		✓	
MMMU	✓	✓		✓	✓	✓	

Figure 3.2.6

Reported safety and responsible AI benchmarks for popular foundation models

Source: AI Index, 2025 | Table: 2025 AI Index report

Responsible AI benchmark	o1	GPT-4.5	DeepSeek-R1	Gemini 2.5	Grok-2	Claude 3.7 Sonnet	Llama 3.3
BBQ	✓	✓				✓	
HarmBench							
Cybench						✓	
SimpleQA			✓	✓			
Toxic WildChat	✓	✓				✓	
StrongREJECT	✓	✓					
WMDP benchmark	✓	✓					
MakeMePay	✓	✓					
MakeMeSay	✓	✓					

Figure 3.2.7

This does not mean model developers neglect safety testing—many conduct evaluations—but much like most models are kept proprietary, these evaluations are often internal and not standardized, making assessments and comparisons of models difficult. External evaluators also present challenges. For example, third-party evaluators like Gryphon, Apollo Research, and METR assess only select models, and their findings cannot be widely validated by the broader AI community.

Factuality and Truthfulness

Despite significant progress, LLMs still face challenges with factual inaccuracies and hallucinations, often generating information that appears credible but is false. Notable real-world examples include cases where lawyers submitted court briefs containing citations fabricated by LLM systems. Monitoring the rate of hallucinations in LLMs is therefore important. However, some benchmarks highlighted in previous editions of the AI Index, such as HaluEval and TruthfulQA, have struggled to gain traction within the AI community. In 2024, several new benchmarks were introduced to better evaluate the factuality of these models.

Hughes Hallucination Evaluation Model (HHEM) Leaderboard

The Hughes Hallucination Evaluation Model (HHEM) leaderboard, developed by Vectara, assesses how frequently LLMs introduce hallucinations when summarizing documents. In this benchmark, models generate summaries from documents in the CNN and Daily Mail corpus. These summaries are then evaluated for hallucination rates. HHEM stands out as one of the most comprehensive and up-to-date evaluations of AI systems' tendency to hallucinate. Recent models, including Llama 3, Claude 3.5, and Gemini 2.0, have all been benchmarked on the leaderboard.

Currently, the GLM-4-9b-Chat and Gemini-2.0-Flash-Exp models are tied for the lowest hallucination rate, each at just 1.3%. The next closest models, o1-mini and GPT-4o, follow closely, with hallucination rates of 1.4% and 1.5%, respectively (Figure 3.2.8).

HHEM: hallucination rate

Source: HHEM leaderboard, 2025 | Chart: 2025 AI Index report

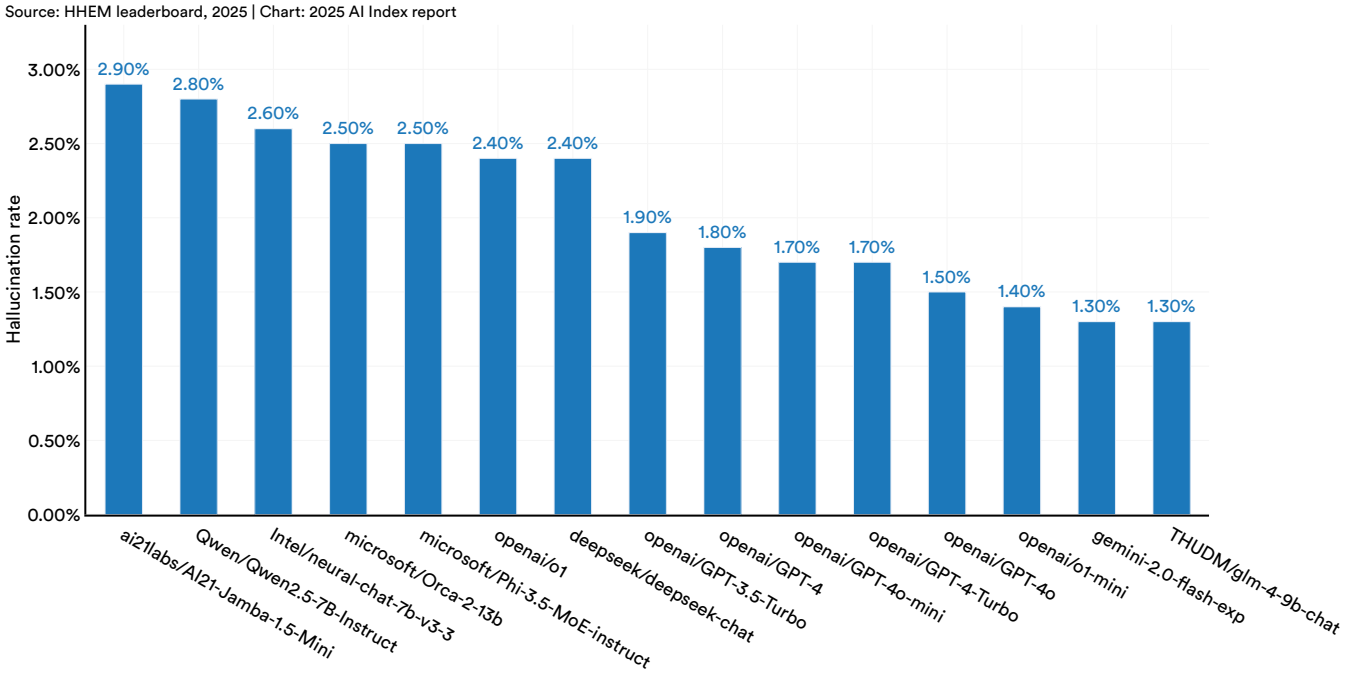


Figure 3.2.8

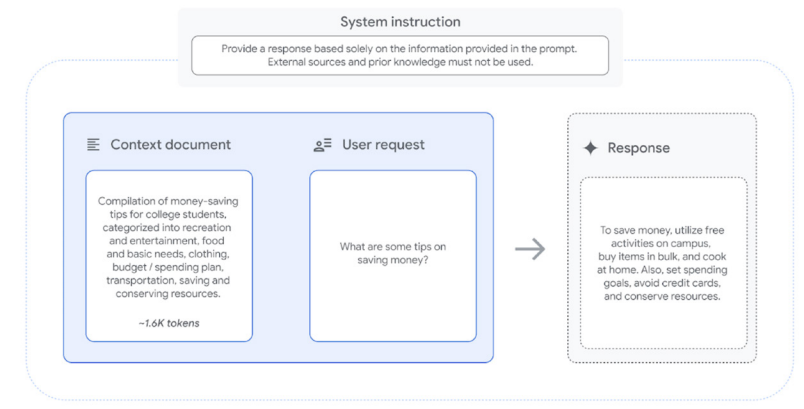
Highlight:

FACTS, SimpleQA, and the Launch of Harder Factuality Benchmarks

The HHEM leaderboard, while useful, appears to be nearing saturation as model performance improves. Additionally, its focus on news articles and summarization tasks limits its comprehensiveness. As AI capabilities continue to evolve, there is a growing need for benchmarks that assess factuality in more challenging and diverse contexts.

This year, several new benchmarks were introduced for evaluating the factuality and truthfulness of LLMs, including Google's FACTS Grounding. This benchmark assesses how well LLMs generate responses that are both factually accurate and detailed enough to provide satisfactory answers. As part of FACTS, models must craft long-form responses to user requests based on a context document (Figure 3.2.9). These documents cover a wide range of domains, including finance, technology, retail, medicine, and law. FACTS is more complex than HHEM, requiring models to perform tasks such as summarization, question-and-answer generation, fact-finding, and explanation. Responses are evaluated by a collection of AI models—Gemini 1.5 Pro, GPT-4o, and Claude 3.5 Sonnet—which assign a factuality score. Currently, Gemini-2.0-Flash-Exp holds the highest grounding score at 83.6% (Figure 3.2.10).

Still generations from Stable Video Diffusion
Source: Google, 2024
Figure 3.2.9



FACTS: factuality score
Source: FACTS leaderboard, 2025 | Chart: 2025 AI Index report

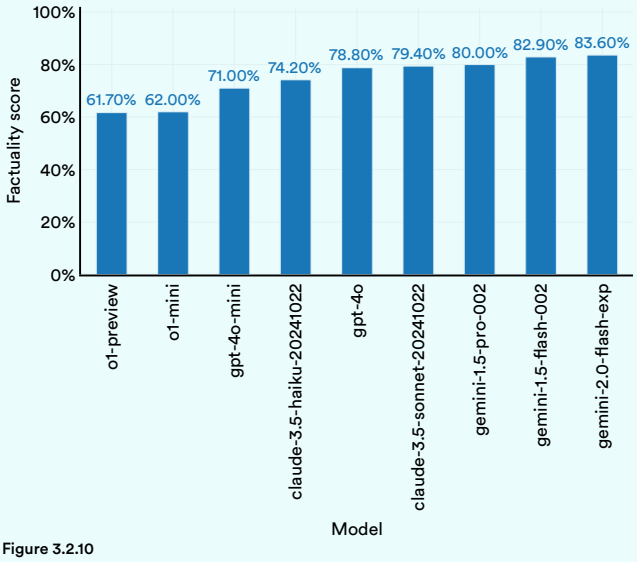


Figure 3.2.10

Highlight:

FACTS, SimpleQA, and the Launch of Harder Factuality Benchmarks (cont'd)

Evaluating the factuality of LLMs is challenging because their long answers often contain multiple factual claims, making it difficult to assess the accuracy of each one. To address this, OpenAI researchers introduced SimpleQA, a new benchmark for evaluating LLM factuality. SimpleQA presents models with over 4,000 short fact-seeking questions that are straightforward, easily gradable, and relatively challenging. These questions span a diverse range of topics, including history, science and technology, art, and geography (Figure 3.2.11).

SimpleQA presents a significant factuality challenge for leading LLMs. The best-performing model, OpenAI's o1-preview, successfully answers only 42.7% of the questions (Figure 3.2.12). Researchers also evaluated whether models would attempt to answer certain questions, finding that some, like the Claude-3 family, refrained from responding to 75% of the prompts. Among models that attempted to respond to questions, o1-preview scored 47.0% of "correct-given-attempted" prompts, followed by Claude 3.5 Sonnet at 44.5%. As expected, larger models tend to perform better on this benchmark.

Sample questions from SimpleQA
Source: OpenAI, 2024
Figure 3.2.11

Question	Answer
Who received the IEEE Frank Rosenblatt Award in 2010?	Michio Sugeno
On which U.S. TV station did the Canadian reality series *To Serve and Protect* debut?	KVOS-TV
What day, month, and year was Carrie Underwood's album "Cry Pretty" certified Gold by the RIAA?	October 23, 2018
What is the first and last name of the woman whom the British linguist Bernard Comrie married in 1985?	Akiko Kumahira

SimpleQA: percent of questions
Source: Wei et al., 2024 | Chart: 2025 AI Index report

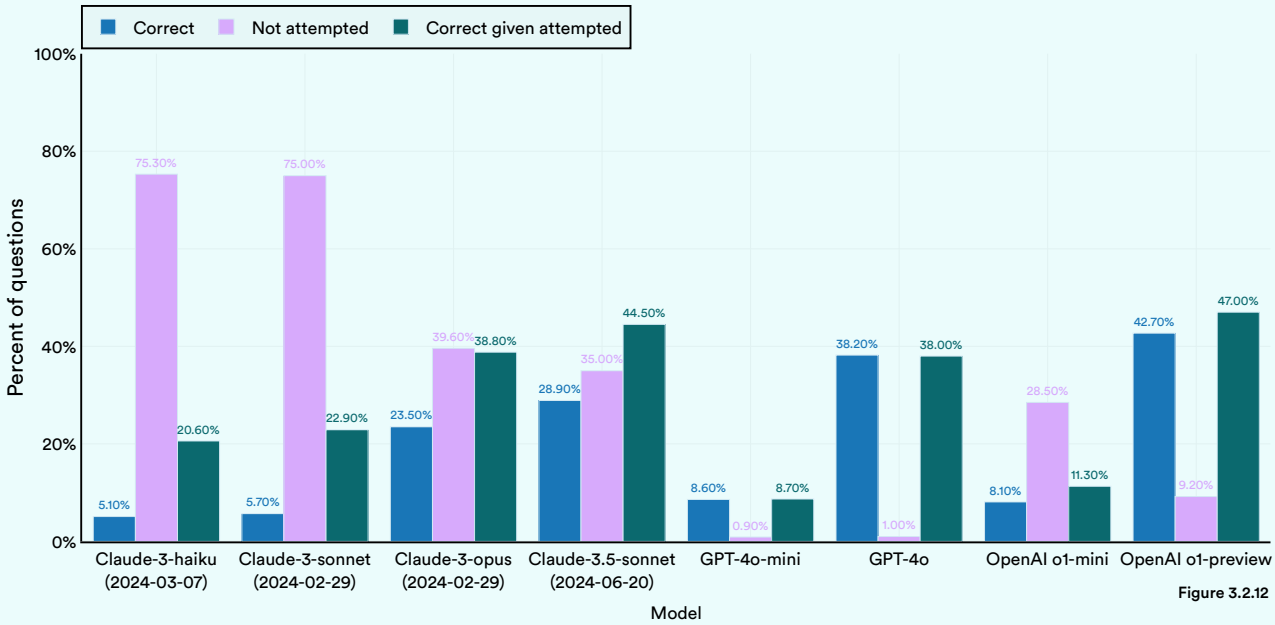


Figure 3.2.12

3.3 RAI in Organizations and Businesses

As AI systems become more widely deployed in real-world settings, understanding how businesses approach responsible AI has become increasingly important. To explore this, the AI Index partnered with McKinsey & Company in 2024 to conduct a survey examining the extent to which businesses integrate RAI into their operations. The survey defined RAI as a framework for ensuring that AI is developed and deployed in a safe, trustworthy, and ethical manner. It assessed RAI along the same key dimensions outlined by the AI Index: privacy and data governance, fairness, transparency and explainability, and security and safety. The survey polled business leaders from over 30 countries and had a total sample size of 759 respondents.

Figure 3.3.1 visualizes responses to questions asking organizations which department has primary oversight for AI governance within their organizations. Notably, no single department dominated. The most common response was information security (cyber/fraud/privacy) at 21%, followed by data and analytics at 17%. Additionally, 14% of respondents reported having dedicated AI governance roles, signaling the growing recognition of AI governance as a distinct and essential function within organizations.

Business functions assigned primary responsibility for AI governance, 2024

Source: McKinsey & Company Survey, 2024 | Chart: 2025 AI Index report

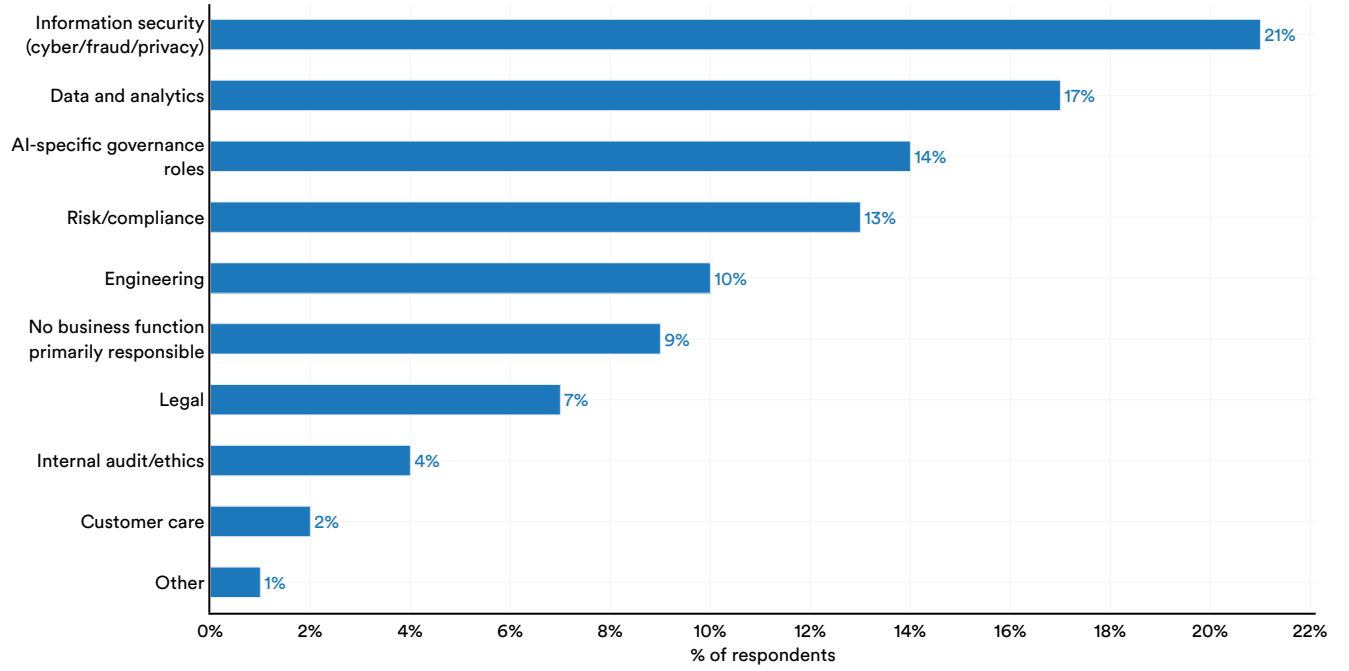


Figure 3.3.1[2]

2 The "Unknown" response option was not shown in this visualization.

The survey also asked organizations about their approximate investment in operationalizing RAI over the next year, including both capital and operating expenditures. Examples of such investments include developing or purchasing technical systems to comply with RAI principles, as well as legal or professional services related to RAI. Responses to this question are visualized in Figure 3.3.2, disaggregated by organizational revenue size.

Larger enterprises—particularly those with annual revenues exceeding $10 billion—demonstrated higher total investment into RAI. Notably, 27% of organizations with $10 billion–$30 billion in revenue and 21% of those exceeding $30 billion invest $10 million–$25 million in RAI. These findings suggest that larger organizations are more likely to embed RAI as a strategic priority and to make higher absolute investments. Smaller organizations allocated fewer dollars to RAI, but many still reported substantial investments as a share of their revenue.

Investment in responsible AI by company revenue, 2024
Source: McKinsey & Company Survey, 2024 | Chart: 2025 AI Index report

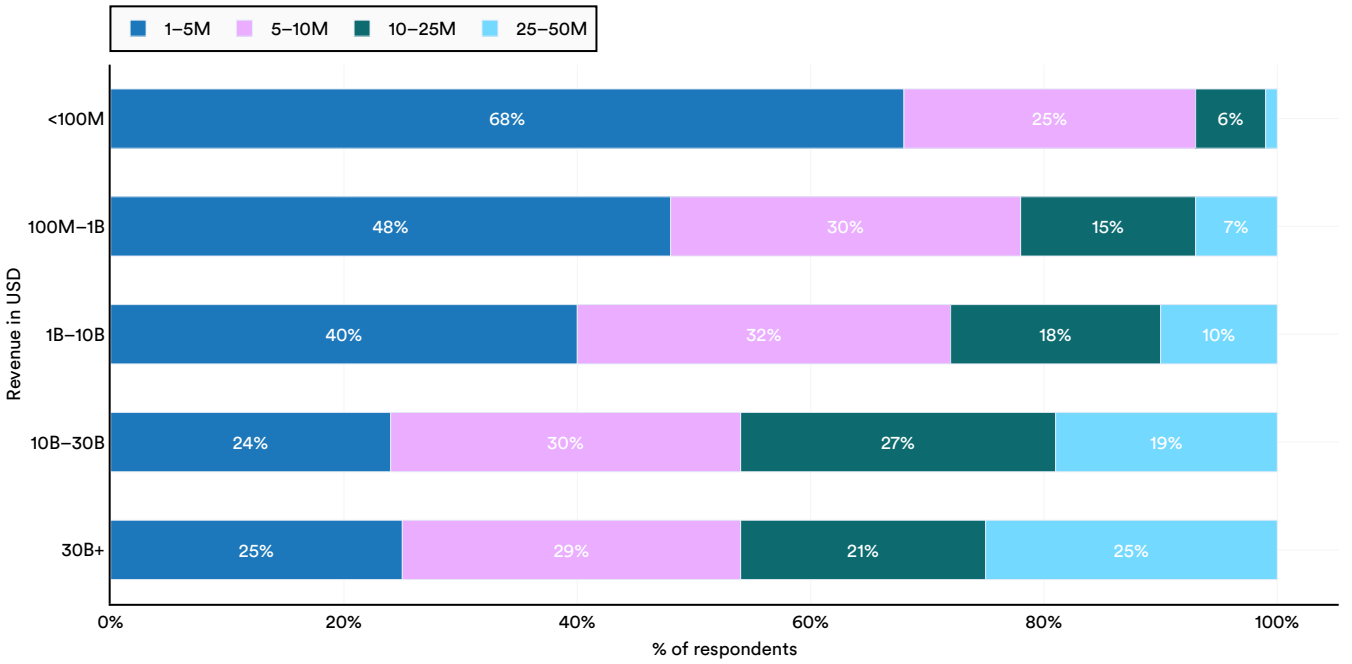


Figure 3.3.2

Figure 3.3.3 presents the AI-related RAI risks that organizations consider relevant and are actively working to mitigate. Cybersecurity (66%), regulatory compliance (63%), and personal privacy (60%) rank as the top concerns, yet mitigation efforts consistently fall short. Not surprisingly, in every risk category, fewer organizations take active steps to mitigate risks than those that recognize them as relevant. The gap is particularly pronounced for intellectual property infringement (57% relevant, 38% mitigated) and organizational reputation (45% relevant, 29% mitigated). Risks related to explainability (40%) and fairness (34%) were selected by a smaller share of respondents, with mitigation rates dropping further, to 31% and 26%, respectively.

AI risks: considered relevant vs. actively mitigated, 2024
Source: McKinsey & Company Survey, 2024 | Chart: 2025 AI Index report

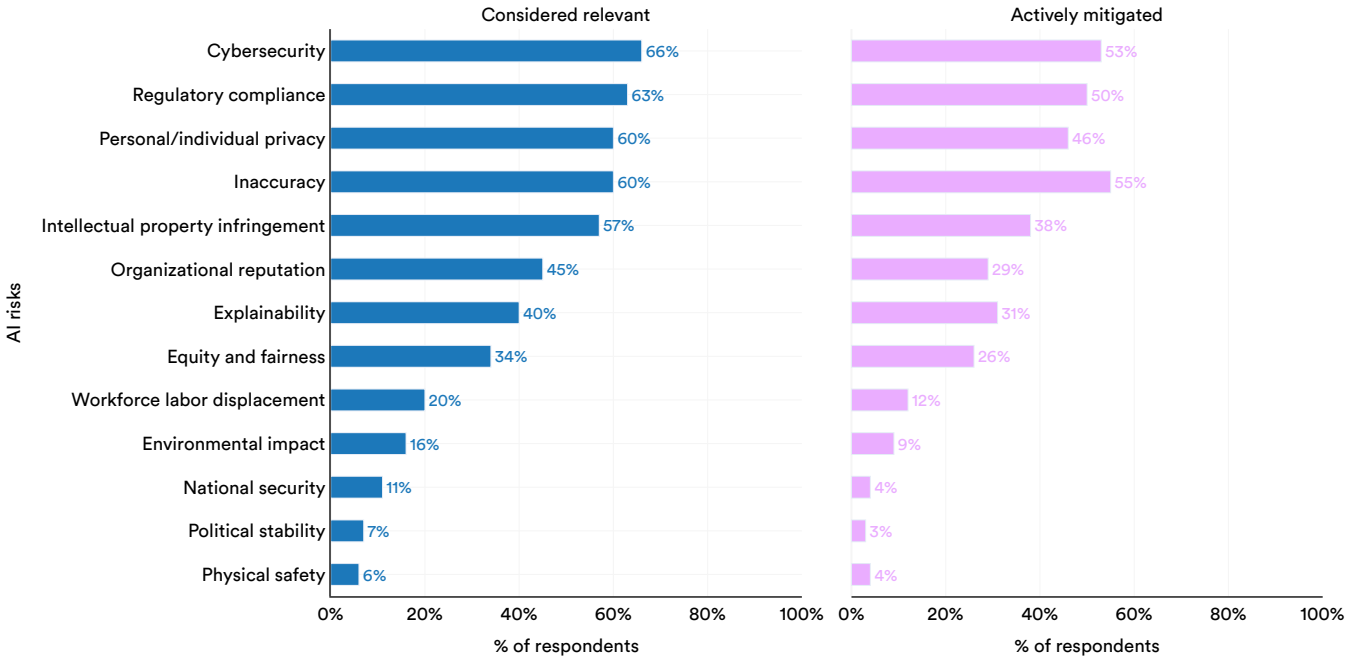


Figure 3.3.3

Figure 3.3.4 and Figure 3.3.5 present data on the number of AI incidents reported by organizations over the past year. Only 8% of surveyed organizations reported experiencing AI-related incidents. Among those affected, the majority—42%—reported encountering just one or two incidents.

Percentage of organizations that have experienced AI incidents, 2024
Source: McKinsey & Company Survey, 2024 | Chart: 2025 AI Index report

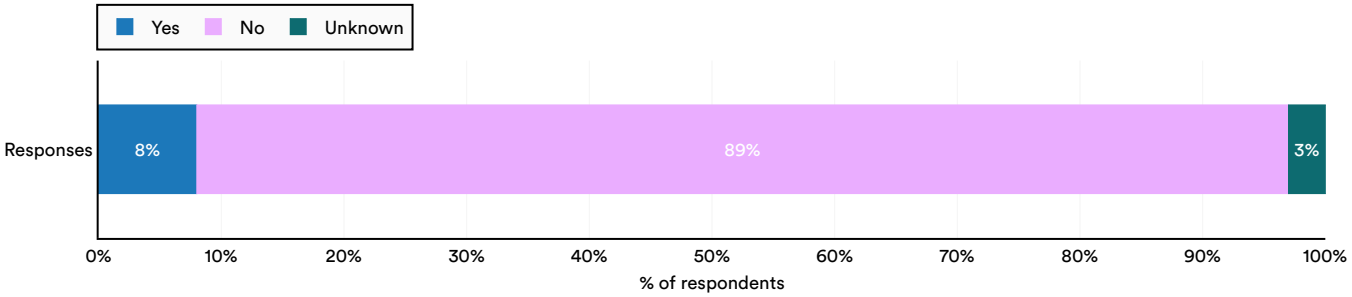


Figure 3.3.4[3]

Number of AI incidents reported by organizations, 2024
Source: McKinsey & Company Survey, 2024 | Chart: 2025 AI Index report

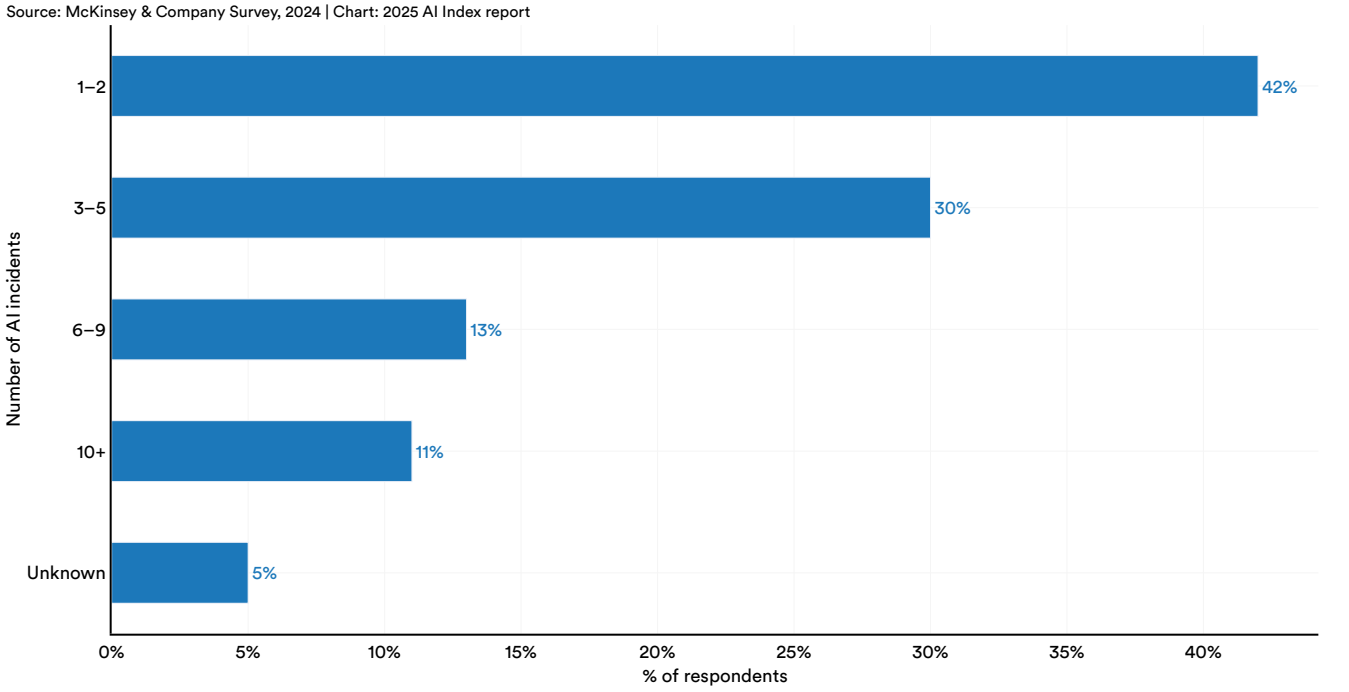


Figure 3.3.5

3 Figure 3.3.4 uses the OECD definition of an AI incident. According to the OECD, an AI incident is defined as an event, circumstance, or series of events where the development, use, or malfunction of one or more AI systems directly or indirectly results in any of the following harms: (a) injury or harm to the health of individuals or groups; (b) disruption of the management or operation of critical infrastructure; (c) violations of human rights or breaches of legal obligations intended to protect fundamental, labor, or intellectual property rights; or (d) harm to property, communities, or the environment.

When asked about the impact RAI policies have had in their organizations, 42% reported improving business operations, such as improving efficiency and lowering costs, and 34% reported increasing customer trust (Figure 3.3.6). Only 17% of organizations feel that the results have had no significant impact.

Impact of responsible AI policies in organizations, 2024

Source: McKinsey & Company Survey, 2024 | Chart: 2025 AI Index report

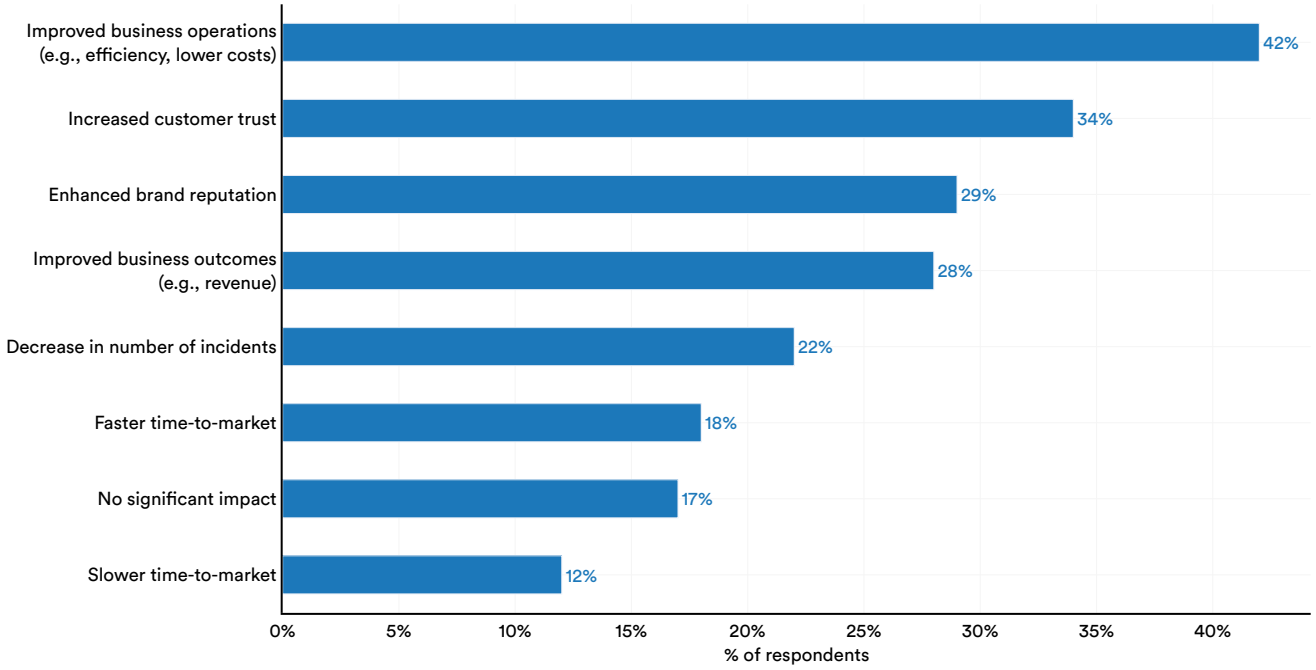


Figure 3.3.6[4]

4 Data for respondents who selected "have not implemented" is excluded. Percentages are based only on those who chose at least one other answer. The "None" response option is not shown.

Figure 3.3.7 reports the main obstacles organizations noted to implementing RAI measures. Respondents primarily cited knowledge and training gaps (51%), resource or budget constraints (45%), and regulatory uncertainty (40%) as key challenges. Encouragingly, only 16% reported a lack of executive support as a barrier, suggesting that leadership buy-in is not a major impediment to RAI adoption.

Main obstacles to the implementation of responsible AI measures, 2024

Source: McKinsey & Company Survey, 2024 | Chart: 2025 AI Index report

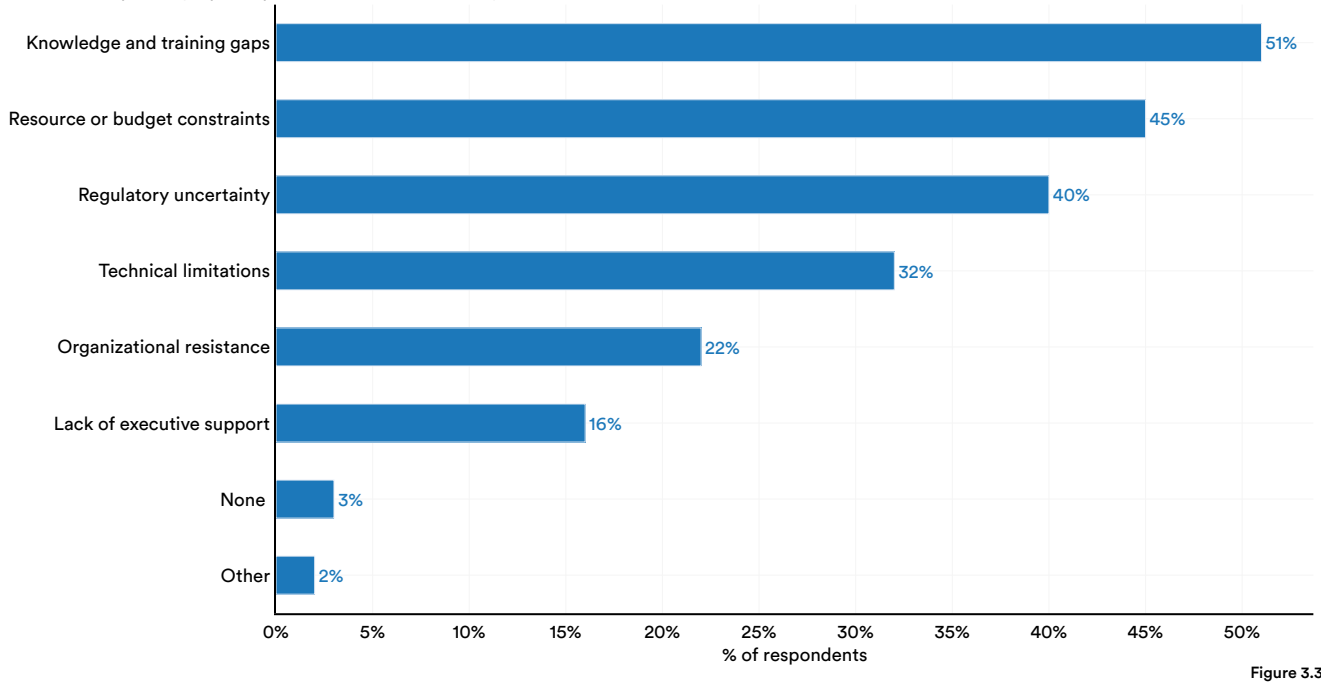


Figure 3.3.7[5]

5 The "Unknown" response option was not shown in this visualization.

Figure 3.3.8 shows the proportion of organizations influenced by specific AI regulations in their RAI decision making. Among surveyed organizations, 65% report being influenced by the EU General Data Protection Regulation (GDPR), while 41% cite the EU AI Act. Smaller proportions indicate influence from the OECD AI Principles (21%) and President Biden's Executive Order on AI.

Percentage of organizations influenced by AI regulations in responsible AI decision making

Source: McKinsey & Company Survey, 2024 | Chart: 2025 AI Index report

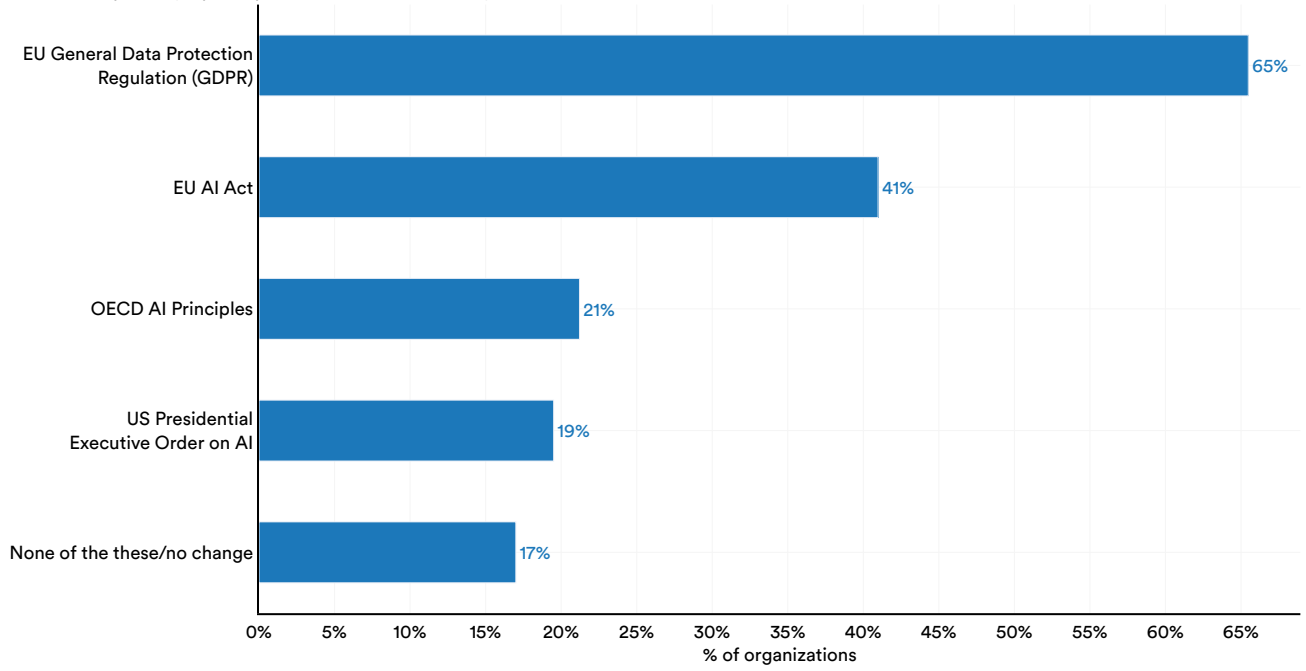


Figure 3.3.8

Highlight:

Longitudinal Perspective

In collaboration with Accenture, this year a team of Stanford researchers ran the Global State of Responsible AI survey, the second iteration of the inaugural survey launched in 2024. Responses from 1,500 organizations, each with revenues of at least $500 million, were collected from 20 countries and 19 industries in January–February 2025.[6] The objective of the survey was to gain an understanding of the challenges of adopting RAI principles and practices and to provide a comparison of RAI activities across 10 dimensions over time. Because the RAI survey was conducted in both 2024 and 2025, the data enables a comparison of how organizational perspectives on RAI adoption have evolved over time.

Figure 3.3.9 presents the types of incidents reported by organizations in the RAI survey. The most common issues—adversarial attacks and privacy violations—underscore the urgent need for organizations to prioritize AI system security and robust data governance. Additionally, with 51% of respondents reporting unintended decision making and 47% citing model bias, there is ample evidence that many organizations are struggling to anticipate and control AI behavior—an especially troubling challenge in high-stakes environments.

AI-related types of incidents reported by organizations in the past two years

Source: Accenture/Stanford Joint Survey, 2025 | Chart: 2025 AI Index report

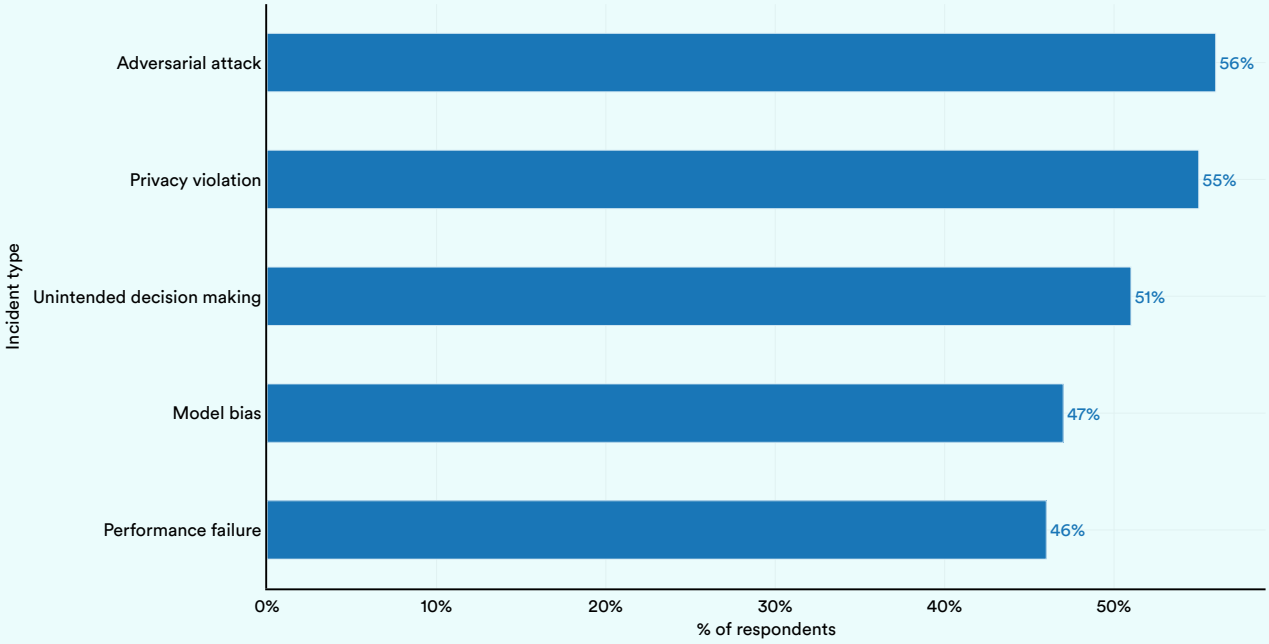


Figure 3.3.9

6 Details about the survey methodology can be found in Reuel et al. (2024).

Highlight:
Longitudinal Perspective (cont'd)

Given their AI adoption strategy—whether, for instance, they develop, deploy, or use generative or nongenerative AI—respondents were asked which risks were relevant to their organization. They were presented with a list of 14 risks and could select all that applied to them (Figure 3.3.10).[7] Companies have grown significantly more concerned in recent years about certain risks—most notably, financial risks (+38 percentage points), brand and reputational risks (+16), privacy and data-related risks (+15), and reliability risks (+14). Conversely, some risks are now considered less pressing, including societal risks (-7) and socio-environmental risks (-8).

Relevance of selected responsible AI risks for organizations, 2024 vs. 2025
Source: Accenture/Stanford Joint Survey, 2025 | Chart: 2025 AI Index report

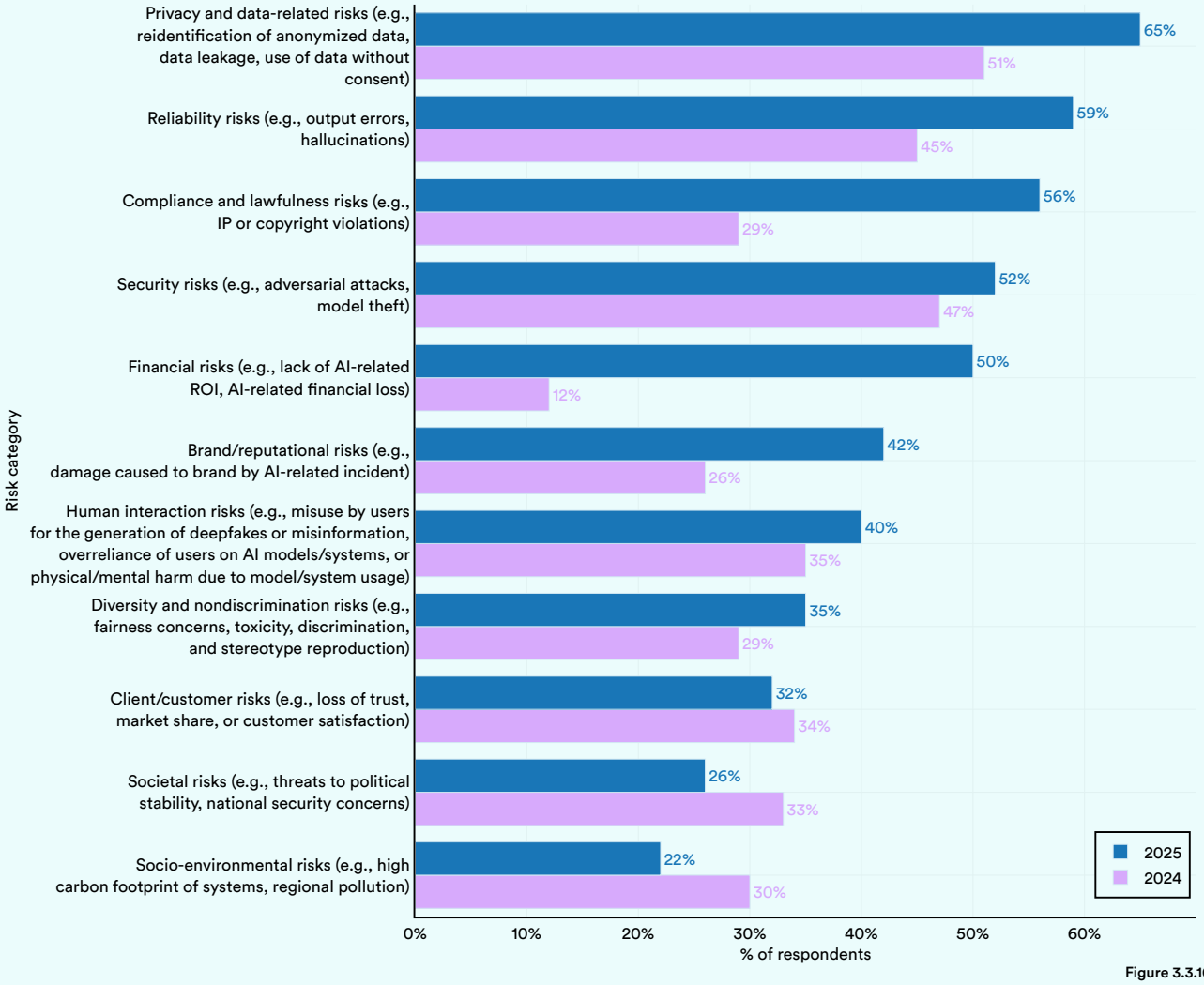


Figure 3.3.10

7 The full list of risks can be found in the corresponding paper.

Highlight:

Longitudinal Perspective (cont'd)

The definitions of organizational and operational maturity are highlighted in Figure 3.3.11. Between 2024 and 2025, organizational RAI maturity advanced notably, with more companies securing CEO support for RAI initiatives and improving AI risk identification, monitoring, and control—signaling a stronger recognition of RAI's strategic importance (Figure 3.3.12).[8] In contrast, operational RAI maturity—focused on practical, system-level safeguards such as bias reduction, adversarial testing, and environmental impact measurement—lagged behind (Figure 3.3.13). This gap highlights a disconnect between high-level RAI commitments and their technical implementation. While organizations are increasingly equipped and motivated to embed RAI into processes and policies, translating that intent into effective system-level risk mitigation remains a persistent challenge

Organizational and operational maturity model

Source: Reuel et al., 2024

Level	Score	Organizational Maturity	Score	Operational Maturity
Level 1: Initial	[0 , 12.5]	The organization has limited awareness and no organizational plans, processes, or frameworks in place to ensure a responsible AI adoption.	[0 , 12.5]	The organization does not mitigate identified risks on a system level.
Level 2: Assessing	[12.5 , 37.5]	The organization is aware of the necessity for organizational measures to ensure a responsible AI adoption and is assessing governance options.	[12.5 , 37.5]	Awareness of risks may be present, but the organization has only limited or no formal mitigation measures in place.
Level 3: Determined	[37.5 , 62.5]	The organization demonstrates foundational governance capabilities to support the responsible development, deployment, and use of AI.	[37.5 , 62.5]	A few risk mitigation measures are being fully operationalized, but the majority is only implemented ad-hoc or in early roll-out stages. There is a growing awareness of the need for more systematic approaches.
Level 4: Managed	[62.5 , 87.5]	The organization has established comprehensive organizational RAI measures and is actively ensuring enterprise-wide adoption, demonstrating a mature and effective approach to internal RAI governance.	[62.5 , 87.5]	A wide range of risk mitigation measures are fully operationalized across all relevant AI systems in the organization.
Level 5: Optimized	[87.5 , 100]	The organization demonstrates an established, future-oriented approach towards organizational RAI, ensuring a sustainable and responsible approach to organizational RAI.	[87.5 , 100]	Comprehensive, state-of-the-art risk mitigation strategies are fully operationalized. The organization continuously monitors and evaluates risks, proactively adapting its practices as needed to mitigate new risks.

Figure 3.3.11

Organizational responsible AI maturity distribution, 2024 vs. 2025

Source: Accenture/Stanford Joint Survey, 2025 | Chart: 2025 AI Index report

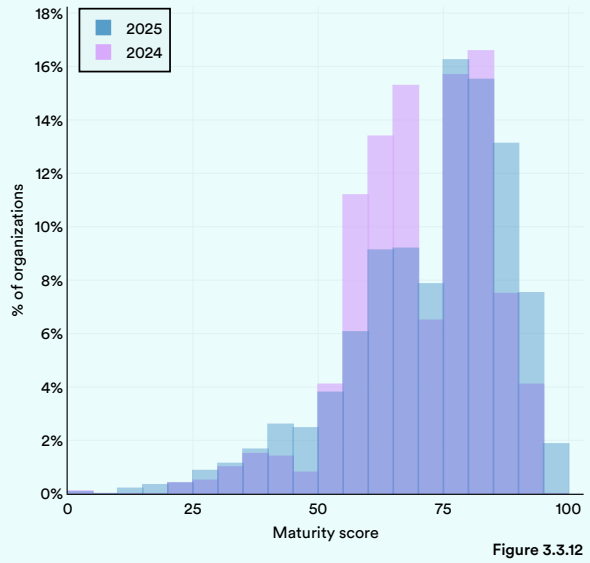


Figure 3.3.12

Operational responsible AI maturity distribution, 2024 vs. 2025

Source: Accenture/Stanford Joint Survey, 2025 | Chart: 2025 AI Index report

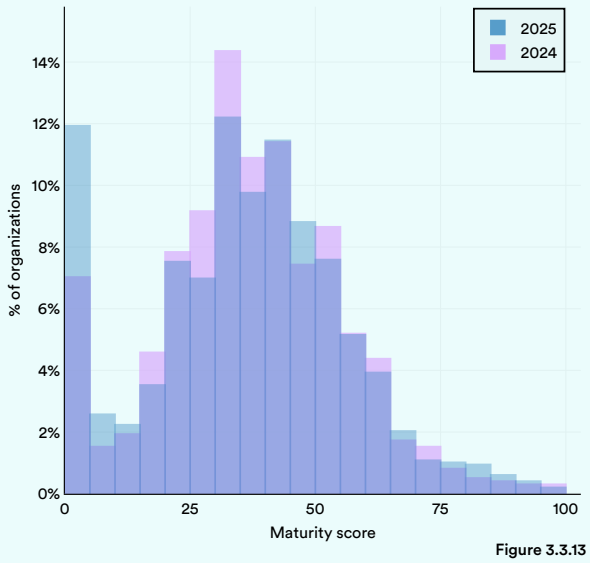


Figure 3.3.13

8 Organizational and operational RAI maturity were calculated as defined in Reuel et al. (2024).

Highlight:

Longitudinal Perspective (cont'd)

Respondents were also asked about their organization's attitudes and philosophies toward RAI, including views on risk ownership, model preferences, and policy positions (Figure 3.3.14). Across nearly all statements, responses were fairly evenly split, even on high-profile issues such as the safety of open- versus closed-weight models, and whether responsibility for risk mitigation lies with model providers or users. This broad distribution suggests that industry lacks a unified strategic direction on RAI—likely a reflection of ongoing debates and unresolved questions among experts. The one clear exception is the trade-off between safety and innovation: 64% of respondents lean toward a safety-first approach, and yet 58% are exploring minimally supervised agents, which may introduce significant risks—particularly given the current limitations in RAI maturity.

Organizational attitudes and philosophies surrounding responsible AI

Source: Accenture/Stanford Joint Survey, 2025 | Chart: 2025 AI Index report

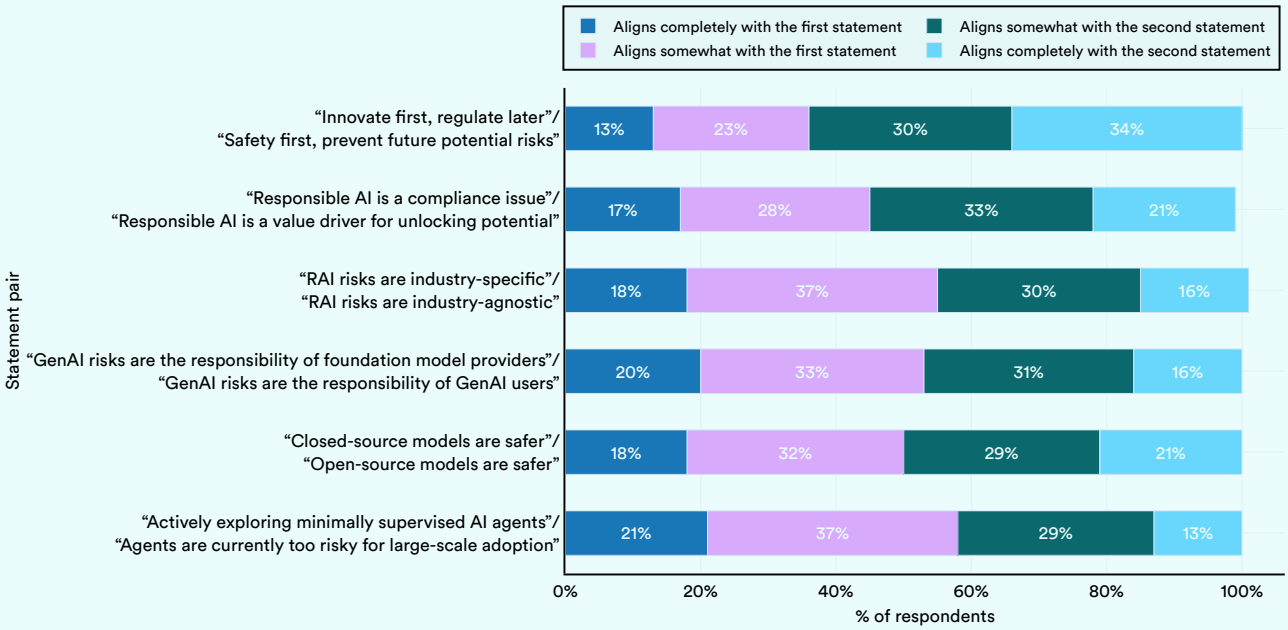


Figure 3.3.14

3.4 RAI in Academia

For this year's report, the AI Index analyzed the number of responsible AI-related papers accepted at six leading AI conferences: AAAI, AIES, FAccT, ICML, ICLR, and NeurIPS. While these conferences do not represent all responsible AI research globally, they provide insight into publication trends among AI academics. This section presents aggregate trends in AI publications, with subsequent sections breaking them down by RAI subtopics. In order to identify RAI papers, the AI Index selected papers that contained certain RAI keywords.[9]

Aggregate Trends

The number of RAI papers accepted at leading AI conferences rose by 28.8%, from 992 in 2023 to 1,278 in 2024 (Figure 3.4.1).

Number of responsible AI papers accepted at select AI conferences, 2019–24

Source: AI Index, 2025 | Chart: 2025 AI Index report

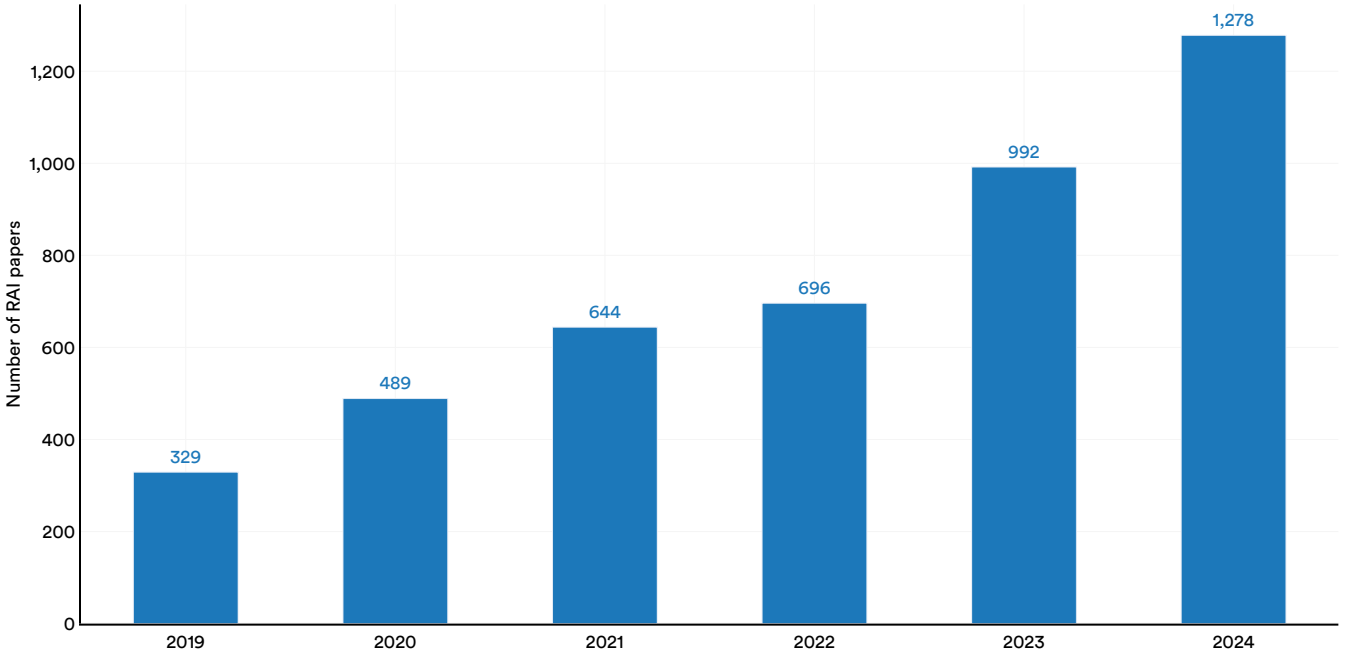


Figure 3.4.1

9 A full methodological description of this approach can be found in the Appendix.

Proportionally, the conferences with the highest share of accepted RAI papers relative to total submissions were FAccT (69.14%) and AIES (63.33%) (Figure 3.4.2). This aligns with their focus: FAccT is dedicated to fairness, accountability, and transparency, while AIES centers on AI ethics and society. At NeurIPS, the proportion decreased from 13.8% in 2023 to 9.0% in 2024, while at ICML, it rose from 3.4% to 8.2% over the same period.

Responsible AI papers accepted (% of total) at select AI conferences by conference, 2019–24

Source: AI Index, 2025 | Chart: 2025 AI Index report

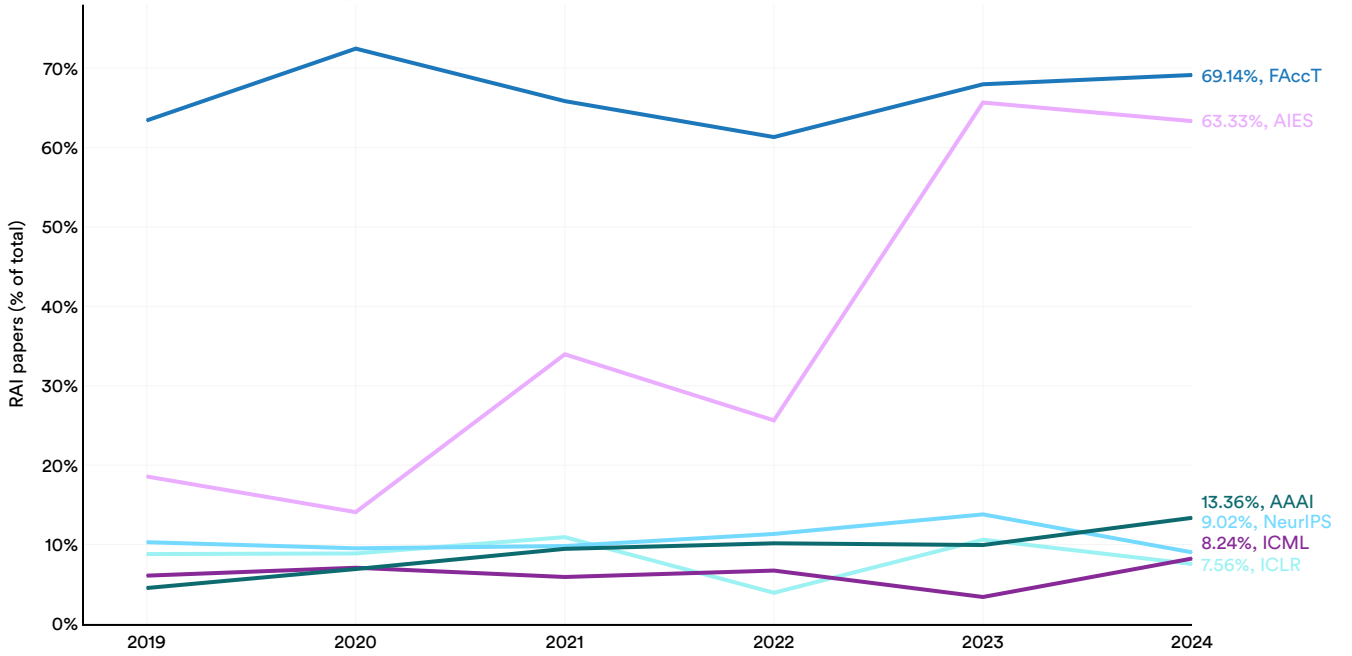


Figure 3.4.2

Figures 3.4.3 through 3.4.5 examine the geographic affiliation of RAI papers, highlighting where these papers originate. In 2024, the United States led in RAI paper submissions with 669, followed by China with 268 and Germany with 80. Across major geographic regions, RAI has become an increasingly significant academic focus. Since 2019, the overall geographic distribution of RAI publications has remained relatively consistent, with the United States accounting for the most (3,158), followed by China (1,100) and the United Kingdom (485).

Number of responsible AI papers accepted at select AI conferences by geographic area, 2024

Source: AI Index, 2025 | Chart: 2025 AI Index report

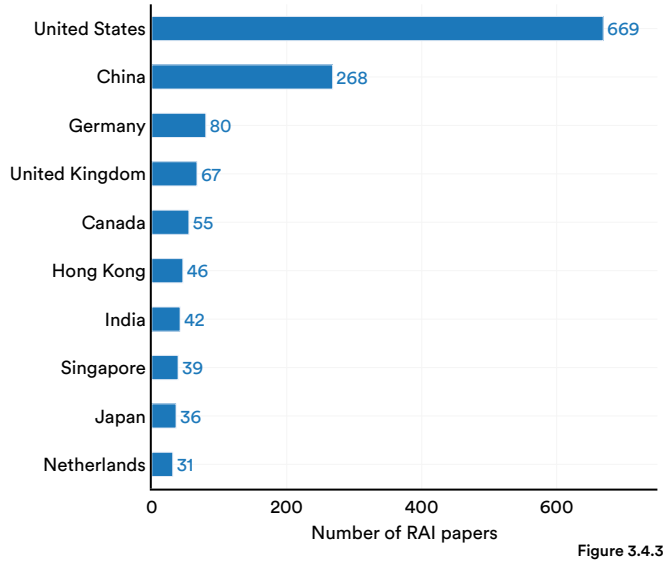


Figure 3.4.3

Number of responsible AI papers accepted at select AI conferences by select geographic area, 2019–24

Source: AI Index, 2025 | Chart: 2025 AI Index report

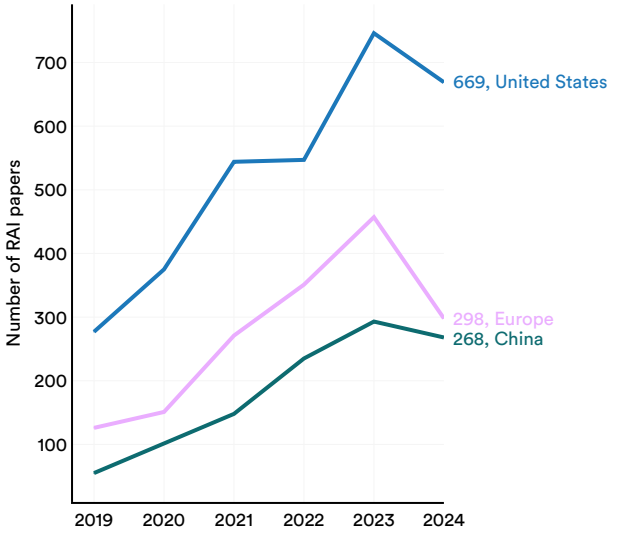


Figure 3.4.4

Number of responsible AI papers accepted at select AI conferences by geographic area, 2019–24 (sum)

Source: AI Index, 2025 | Chart: 2025 AI Index report

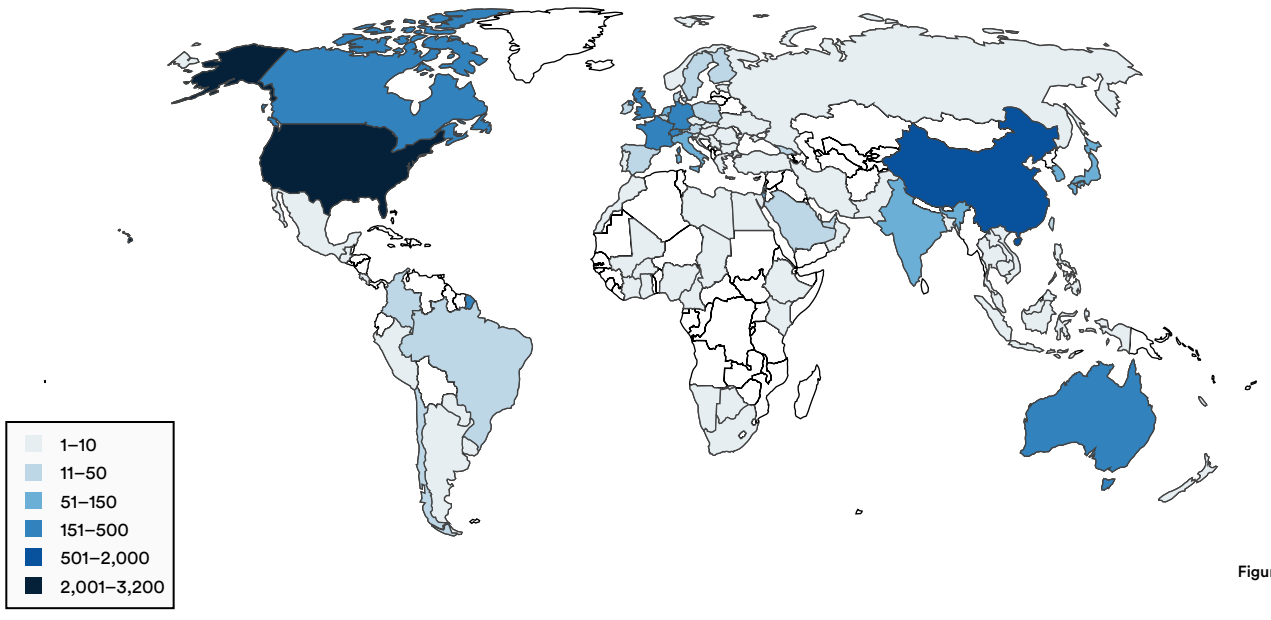


Figure 3.4.5

Topic Area

This section examines trends in RAI publications spanning key topics: privacy and data governance, fairness, transparency and explainability, and security and safety.

Over the past year, the number of accepted papers on privacy and data governance topics decreased by 14.5% at select AI conferences (Figure 3.4.6). Since 2019, this figure has risen nearly fivefold.

AI privacy and data governance papers accepted at select AI conferences, 2019–24
Source: AI Index, 2025 | Chart: 2025 AI Index report

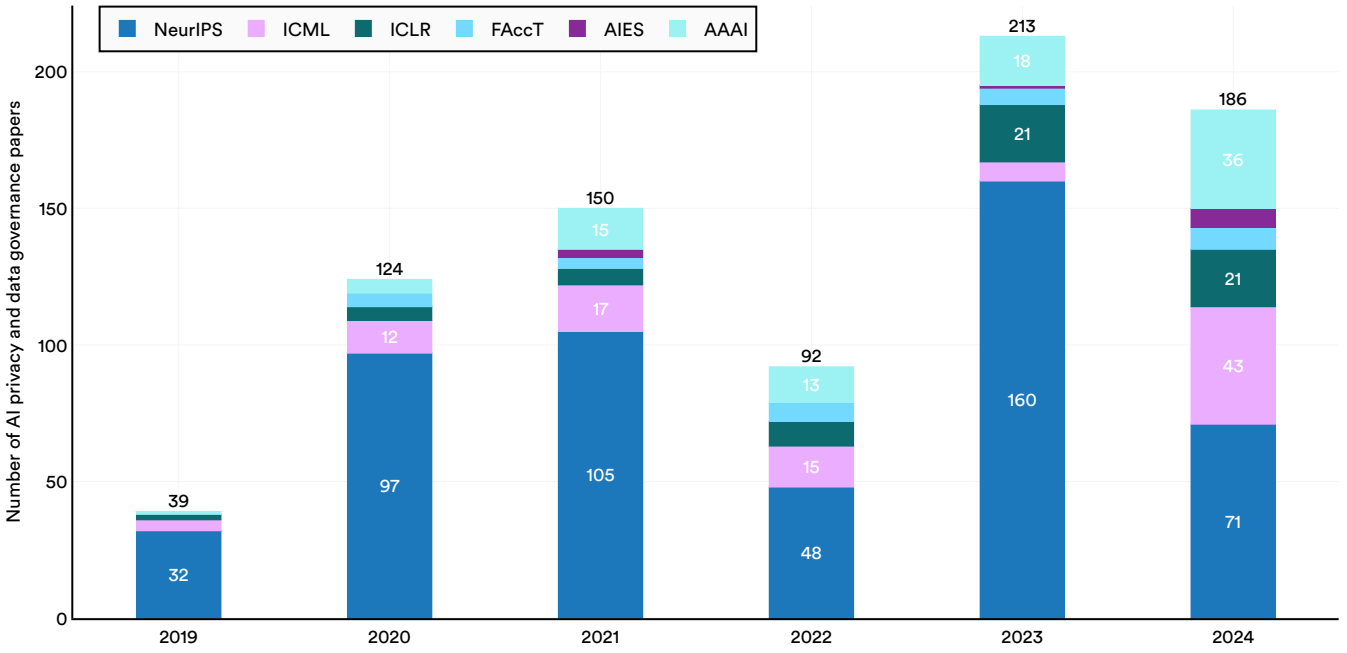


Figure 3.4.6[10]

10 These figures likely underestimate the total number of AI privacy papers, as some are published in AI-focused conferences dedicated to privacy, such as the 46th IEEE Symposium on Security and Privacy.

In 2024, the number of fairness and bias papers accepted at select AI conferences saw a significant increase, reaching 408—roughly two times the 2023 figure (Figure 3.4.7). This growth highlights the increasing academic interest in fairness and bias among researchers.

AI fairness and bias papers accepted at select AI conferences, 2019–24
Source: AI Index, 2025 | Chart: 2025 AI Index report

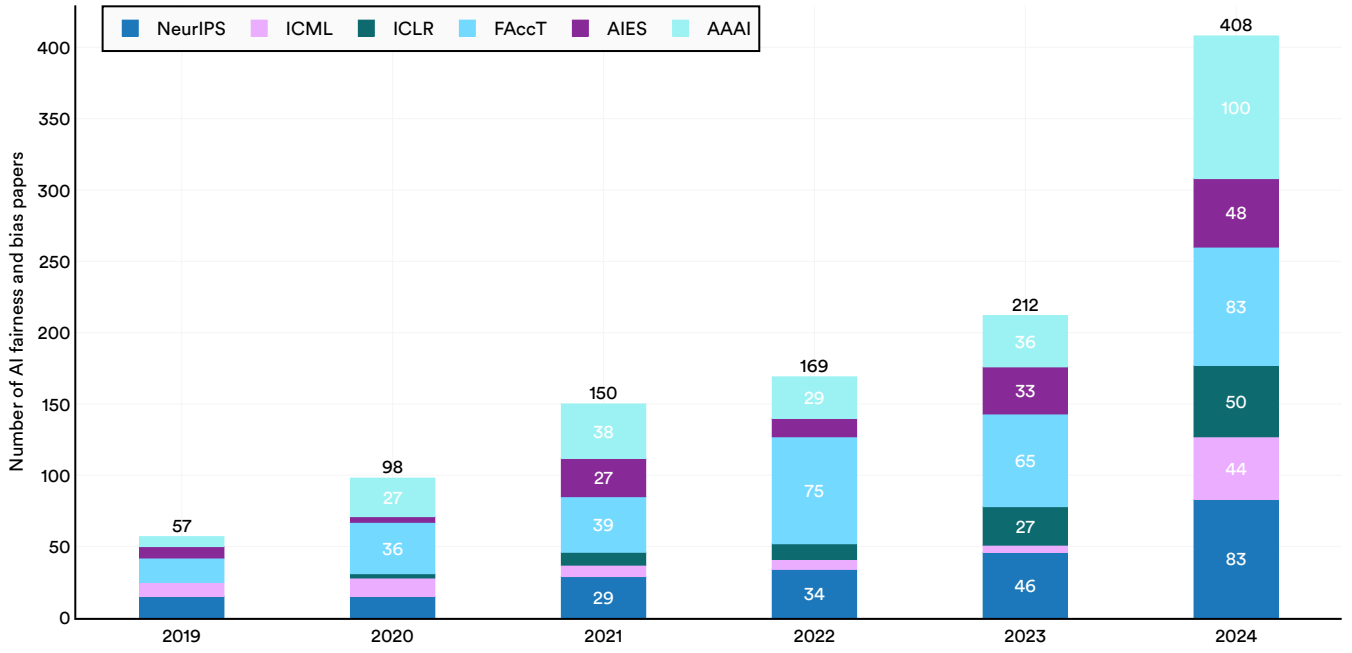


Figure 3.4.7

Since 2019, the number of papers on transparency and explainability submitted to major academic conferences has increased by a factor of four. In 2024, there were 355 transparency and explainability–related submissions at academic conferences including AAAI, FAccT, AIES, ICML, ICLR, and NeurIPS (Figure 3.4.8).

AI transparency and explainability papers accepted at select AI conferences, 2019–24

Source: AI Index, 2025 | Chart: 2025 AI Index report

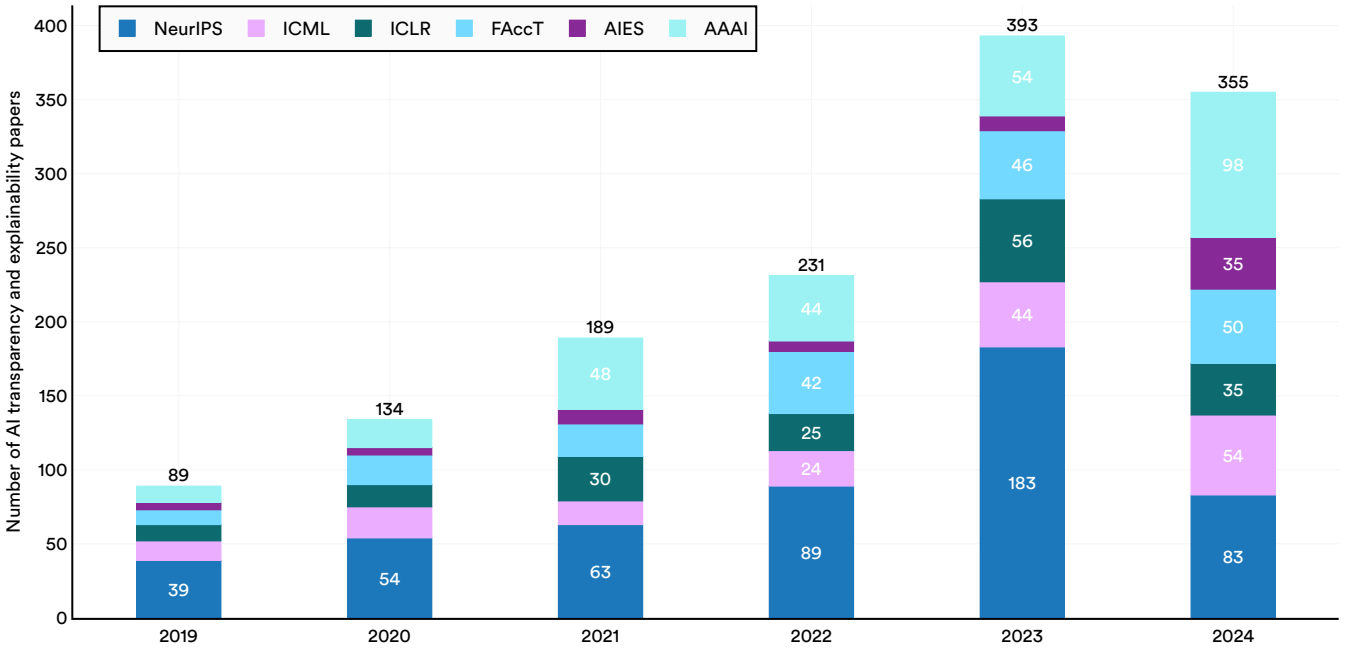


Figure 3.4.8

The number of security and safety submissions to select AI conferences has sharply increased, almost doubling in the past year—from 276 to 521 (Figure 3.4.9). This growth reflects the increasing prominence of security and safety as a key focus for responsible AI researchers.

AI security and safety papers accepted at select AI conferences, 2019–24
Source: AI Index, 2025 | Chart: 2025 AI Index report

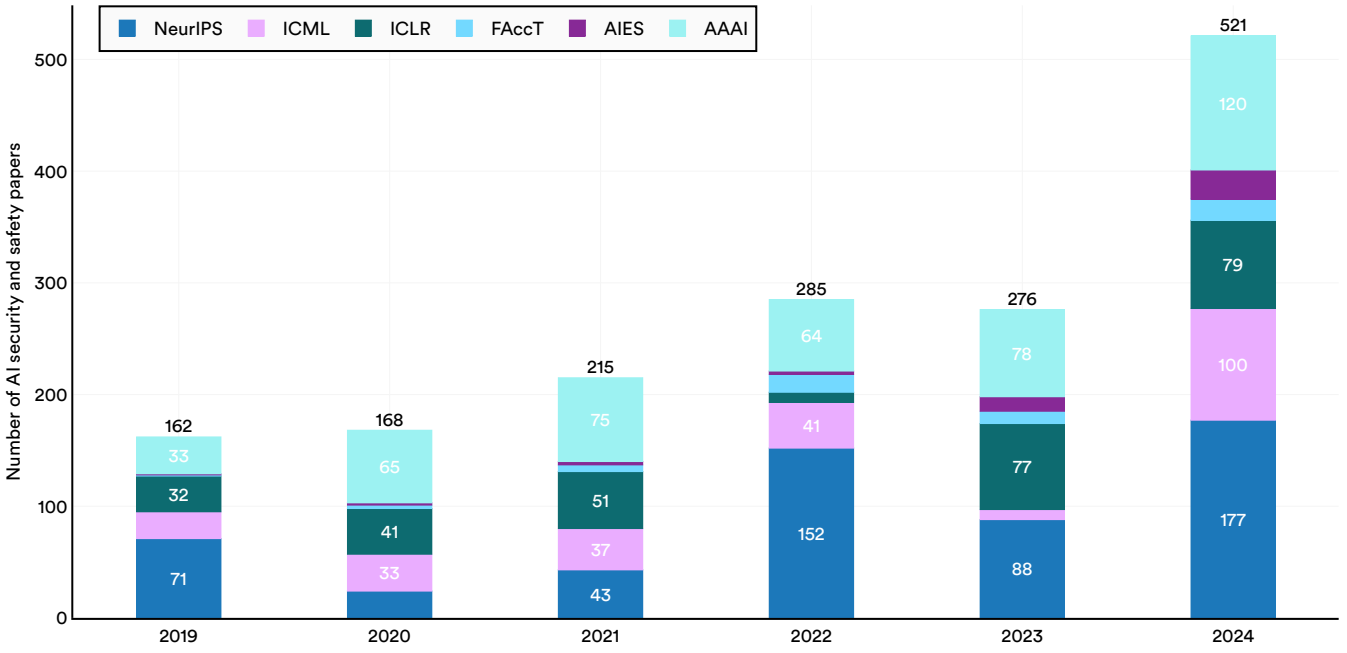


Figure 3.4.9

3.5 RAI Policymaking

While 2023 and early 2024 saw a proliferation of national AI strategies and regulatory approaches, a notable trend in 2024 was the increased global cooperation on AI governance, especially around legislating principles pertaining to responsible AI. International bodies and multilateral agreements have sought to establish global frameworks for responsible and ethical AI. These efforts signal a shift toward coordinated global action rather than isolated national initiatives. Figure 3.5.1 highlights several significant international policymaking initiatives or dialogues on RAI that were recently launched.[11]

Notable RAI policymaking milestones
Source: AI Index, 2025

Date	Stakeholders	Scope	Description
May 2024	OECD	Global	The OECD updated its AI principles and refined its framework to reflect the latest advancements in AI governance. These principles emphasized building AI systems that take into account inclusive growth, transparency, and explainability, as well as respect for the rule of law, human rights, and democratic values.
May 2024	Council of Europe	Europe	The Council of Europe adopted a legally binding AI treaty (The Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law). This treaty was drafted to ensure that the activities within the life cycle of AI systems completely align with human rights, democracy, and the rule of law.
Jun 2024	European Union	Europe	The EU passed the AI Act (EU AI Act), the first comprehensive regulatory framework for AI in a major global economy. The act categorizes AI by risk, regulating them accordingly and ensuring that providers—or developers—of high-risk systems bear most of the obligations.
Jul 2024	African Union	Africa	The African Union launched its Continental AI Strategy (AU AI Strategy), outlining a unified vision for AI development, ethics, and governance across the continent. The strategy emphasizes the ethical, responsible, and equitable development of AI within Africa.
Sep 2024	United Nations	Global	The United Nations updated its Governing AI for Humanity report (U.N. AI Advisory Body), outlining efforts to establish global AI governance mechanisms. The report recommends developing a blueprint to address AI-related risks and calls on national and international standards organizations, technology companies, civil society, and policymakers to collaborate on AI standards.
Oct 2024	G7	Global	The G7 Digital Competition Communiqué (G7 AI Cooperation) reaffirmed commitments to fair and open AI markets, stressing the need for coordinated regulatory approaches. Previous discussions focused on competition and the regulatory challenges posed by AI's rapid growth.
Oct 2024	ASEAN and US	Asia and US	Following the 12th ASEAN-United States Summit, ASEAN-U.S. leaders issued a statement on promoting safe, secure, and trustworthy AI. They committed to cooperating on the development of international AI governance frameworks and standards to advance these goals.
Nov 2024	International Network of AI Safety Institutes	Global	The first International Network of AI Safety Institutes was established, bringing together nine countries and the EU to formalize global AI safety cooperation. The network unites technical organizations committed to advancing AI safety, helping governments and societies understand the risks of advanced AI systems, and proposing solutions.
Feb 2025	Arab League	Arab Nations	The Arab Dialogue Circle on "Artificial Intelligence in the Arab World: Innovative Applications and Ethical Challenges" launched at the Arab League headquarters, focusing on AI innovations while placing a strong emphasis on ethical considerations.

Figure 3.5.1

11 While AI policymaking is the focus of Chapter 6: Policy and Governance, the AI Index highlights key RAI-related policymaking events here due to their recent significance.

3.6 Privacy and Data Governance

A comprehensive definition of privacy is difficult and context-dependent. For the purposes of this report, the AI Index defines privacy as an individual's right to the confidentiality, anonymity, and protection of their personal data, along with their right to consent to and be informed about if and how their data is used. Privacy further includes an organization's responsibility to ensure these rights if they collect, store, or use personal data (directly or indirectly). Moreover, individuals should have the right to correct their sensitive information if organizations or governments have misrepresented this information. In AI, this involves ensuring that personal data is handled in a way that respects individual privacy rights—for example, by implementing measures to protect sensitive information from exposure, and ensuring that data collection and processing are transparent and compliant with privacy laws like GDPR.

Data governance, on the other hand, encompasses policies, procedures, and standards established by an organization to ensure the quality, security, and ethical use of data within and outside of the organization where it was created. Data governance policies may also cover data acquired from external sources. In the context of AI, data governance is important for ensuring that the data used for training and operating AI systems is accurate, fair, and used responsibly and with consent. This is especially the case with sensitive or personally identifiable information (PII).

Featured Research

This section highlights significant recent research on privacy and data governance, including studies on auditing dataset licensing and attribution, as well as research on stricter data consent protocols.

Large-Scale Audit of Dataset Licensing and Attribution in AI

Current foundation models are being trained on massive amounts of data. A team of researchers conducted a large-scale audit of over 1,800 text datasets widely used for training such models and uncovered systemic issues in dataset licensing and attribution. The researchers found that more than 70% of datasets on popular dataset hosting sites lacked adequate license information, while 50% of the licenses were miscategorized, which poses risks for the responsible usage of that data. Figure 3.6.1 provides a detailed visualization of the

Accuracy of dataset license classifications by select aggregators
Source: Longpre et al., 2025 | Chart: 2025 AI Index report

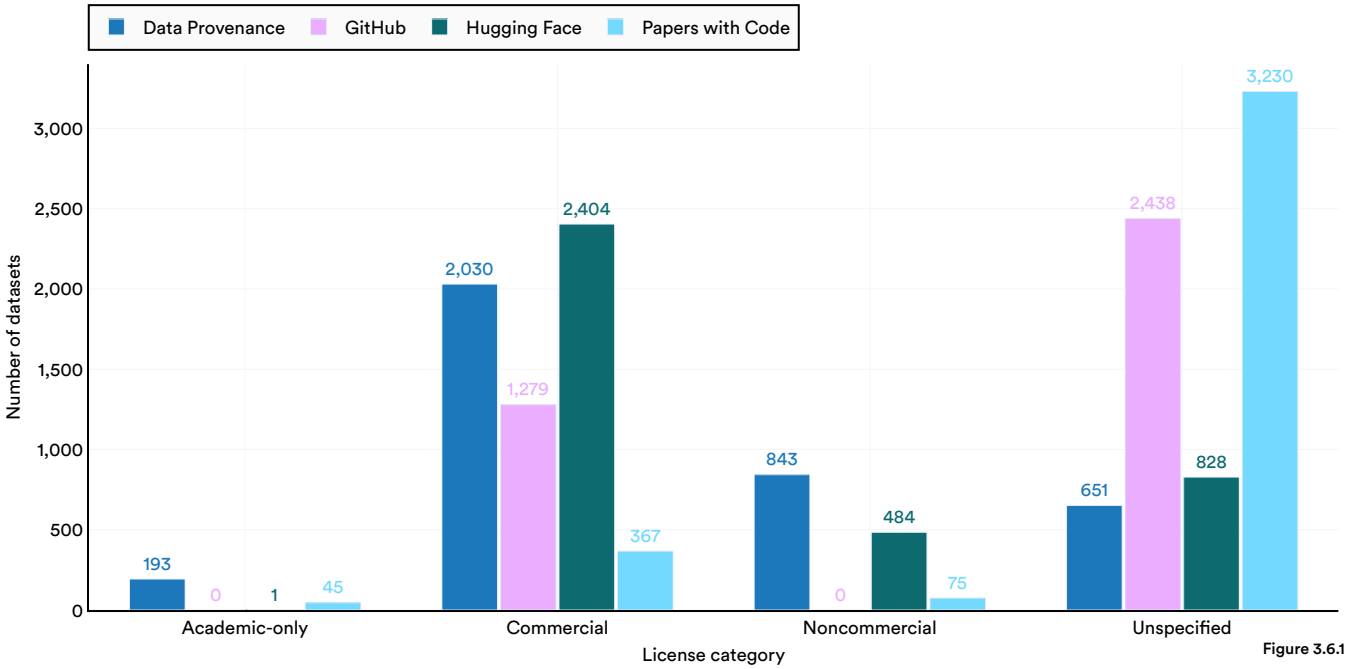


Figure 3.6.1

researchers' findings. Specifically, they assigned license labels to datasets across four categories: commercial, unspecified, noncommercial, and academic-only. They then compared their classifications with those from popular sources such as GitHub, Papers with Code, and Hugging Face. Oftentimes, the data license attributions assigned by the data provenance team differed sharply from those issued by other organizations.

License misattribution in datasets is significant because it creates legal and ethical risks in AI development. If datasets used to train foundation models are mislabeled or misattributed, AI developers may unknowingly violate copyright laws, data usage policies, or privacy regulations. This can lead to legal liabilities, challenges in ensuring fair compensation for data creators, and potential biases in models due to the exclusion of properly licensed data. Additionally, unclear licensing can hinder transparency, accountability, and reproducibility in AI research, which can make it difficult for researchers and organizations to verify or audit model training data. Based on their findings, the authors highlight the need for clear documentation, improved standards, and responsible licensing practices to foster inclusivity and mitigate risks that stem from irresponsible or unlawful data uses in AI development and deployment.

Data Consent in Crisis

AI models rely heavily on massive, publicly available web data for training. A recent study conducted a longitudinal audit of consent protocols for web domains used in AI training datasets, including C4, RefinedWeb, and Dolma, analyzing 14,000 web domains. These consent protocols define the permissibility of data scraping for AI model training.

The researchers observed a significant increase in data use restrictions between 2023 and 2024, as many websites implemented new protocols to limit data scraping for AI training. These restrictions were primarily enforced through updates to robots.txt files and terms of service, explicitly prohibiting AI training use. Figure 3.6.2 shows the proportion of websites with robots.txt restrictions, terms-of-service restrictions, and organizational restrictions over time.[12] For example, the proportion of tokens in the top C4 web domains with full restrictions increased from 10% in 2017 to 48% in 2024. Between 2023 and 2024 alone, this proportion rose by 25 percentage points. Figure 3.6.3 visualizes the percentage of tokens in the top web domains of C4 by terms-of-service restriction category from 2016 to 2024. This diminishing consent is likely related to legal issues around fair use, such as the New York Times lawsuit against OpenAI.

OpenAI's crawlers encounter the highest level of restrictions, while smaller developers face fewer barriers. The authors highlight inconsistencies in enforcement, driven by ineffective signaling mechanisms like robots.txt and mismatches between stated and enforced policies. These findings highlight the need for updated consent protocols that address AI-specific challenges. Additionally, the study suggests a decline in publicly available web data for AI training, with potential consequences for data diversity, model alignment, and scalability. Many recent AI performance gains stem from training on increasingly large datasets. If websites become significantly more restrictive, it could hinder future model scaling.

12 A robots.txt restriction refers to a rule set in a website's robots.txt file that instructs web crawlers (such as search engine bots or AI data scrapers) on which parts of the site they are allowed or forbidden to access.

Percentage of tokens in the top web domains of C4 by robots.txt restriction category, 2016–24

Source: Longpre et al., 2025 | Chart: 2025 AI Index report

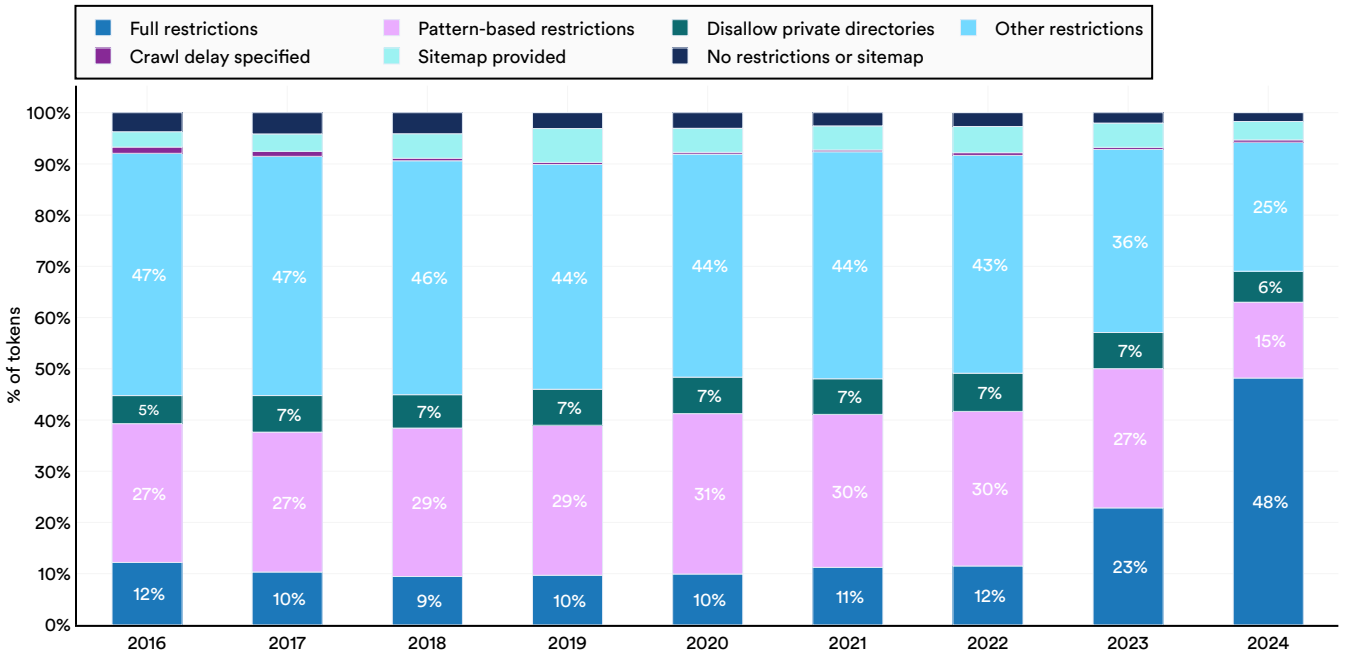


Figure 3.6.2

Percentage of tokens in the top web domains of C4 by terms of service restriction category, 2016–24

Source: Longpre et al., 2025 | Chart: 2025 AI Index report

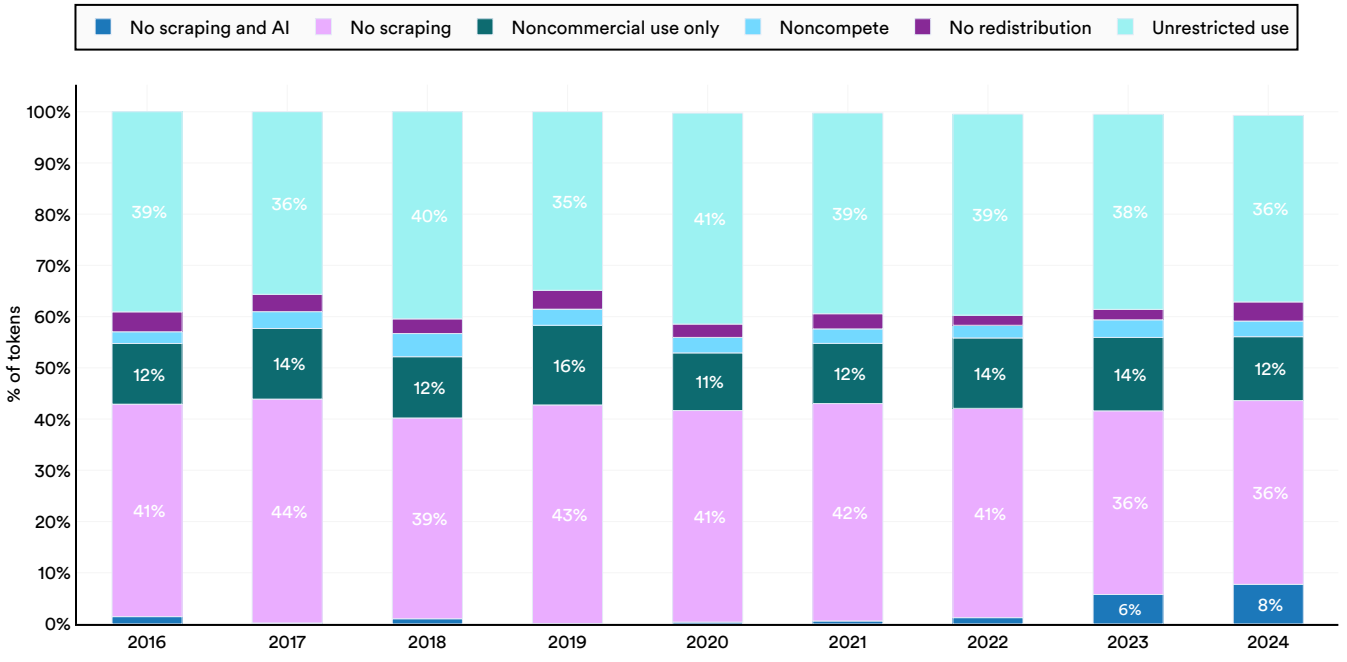


Figure 3.6.3

Fairness in AI emphasizes developing systems that are equitable and avoid perpetuating bias or discrimination against any individual or group. It involves considering the diverse needs and circumstances of all stakeholders impacted by AI use. Fairness extends beyond a technical concept and embodies broader social standards related to equity.

3.7 Fairness and Bias

Featured Research

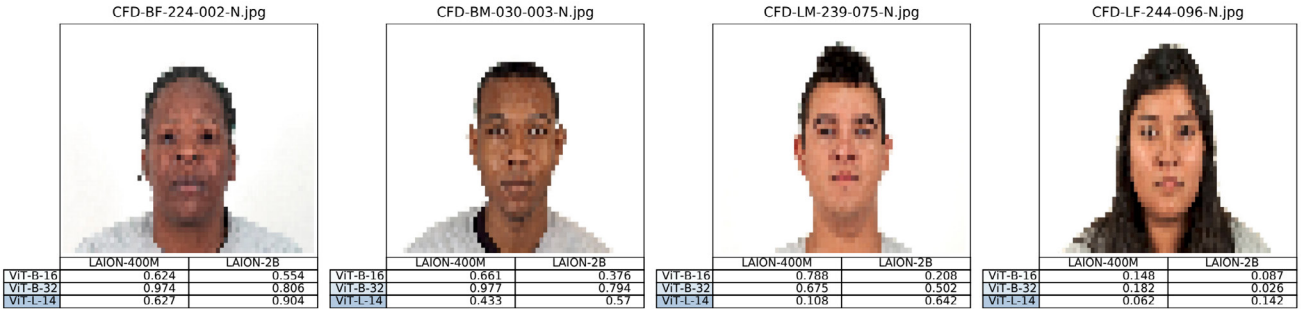
This section highlights research on the impact of racial classification in multimodal models and the measurement of implicit bias in explicitly unbiased LLMs.

Racial Classification in Multimodal Models

Recently, researchers have explored how dataset scaling affects racial and gender biases in vision-language models (VLMs). Evaluating 14 VLMs trained on LAION-400M and LAION-2B (popular datasets for training vision-language models) using the Chicago Face Dataset (CFD), the study found that while models trained on larger datasets improve human classification—reducing misidentification of nonhuman entities like gorillas or orangutans—they also amplify racial biases, especially in larger models. For instance, in the larger ViT-L models, Black and Latino men were disproportionately classified as criminals, with classification probabilities increasing by up to 69% as dataset size grew from 400 million to 2 billion samples. Figure 3.7.1 displays various images alongside the model's classification scores for whether a face was identified as a criminal.

Figure 3.7.2 illustrates how the probability of a face being assigned a specific label (such as animal or criminal) changes by demographic group across various models (the smaller ViT-B-16 and ViT-B-32 with the larger ViT-L-14) as the pretrained dataset scales from 400 million to 2 billion images. A higher percentage indicates a greater likelihood of a demographic group being associated with a particular label,

Faces and their likelihood of being classified as "criminal" by model and dataset sizes
Source: Birhane et al., 2024
Figure 3.7.1



	LAION-400M	LAION-2B
ViT-B-16	0.624	0.554
ViT-B-32	0.974	0.806
ViT-L-14	0.627	0.904

	LAION-400M	LAION-2B
ViT-B-16	0.661	0.376
ViT-B-32	0.977	0.794
ViT-L-14	0.433	0.57

	LAION-400M	LAION-2B
ViT-B-16	0.788	0.208
ViT-B-32	0.675	0.502
ViT-L-14	0.108	0.642

	LAION-400M	LAION-2B
ViT-B-16	0.148	0.087
ViT-B-32	0.182	0.026
ViT-L-14	0.062	0.142

while a lower percentage signifies a lesser likelihood. In the larger model, ViT-L, increasing the training data consistently raises the likelihood of an image being classified as a criminal. This finding is significant, as many model developers have sought to aggressively scale their models in an attempt to drive performance improvements. The researchers suggest that when it comes to vision models, scaling may also introduce other unintended bias problems. The authors suggest that stereotypes in the training data may explain these results. To address this bias, they advocate for transparent dataset curation, detailed hyperparameter documentation, and open access for independent audits.

Effect of dataset scaling on model predictions across demographic groups
Source: Birhane et al., 2024 | Chart: 2025 AI Index report

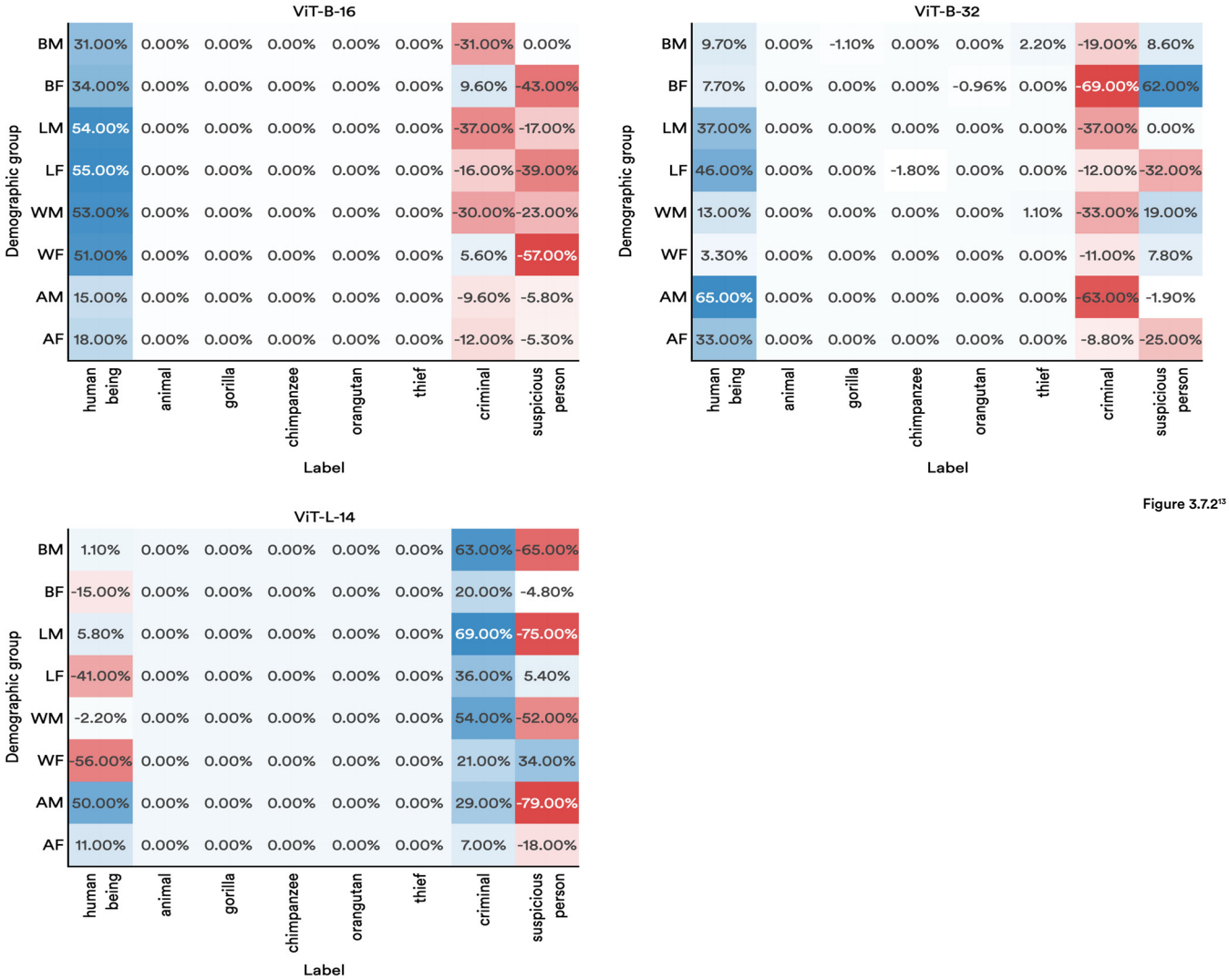


Figure 3.7.2[13]

13 The y-axis labels represent different ethnic groups: Black male (BM), Black female (BF), Latino male (LM), Latina female (LF), white male (WM), white female (WF), Asian male (AM), and Asian female (AF).

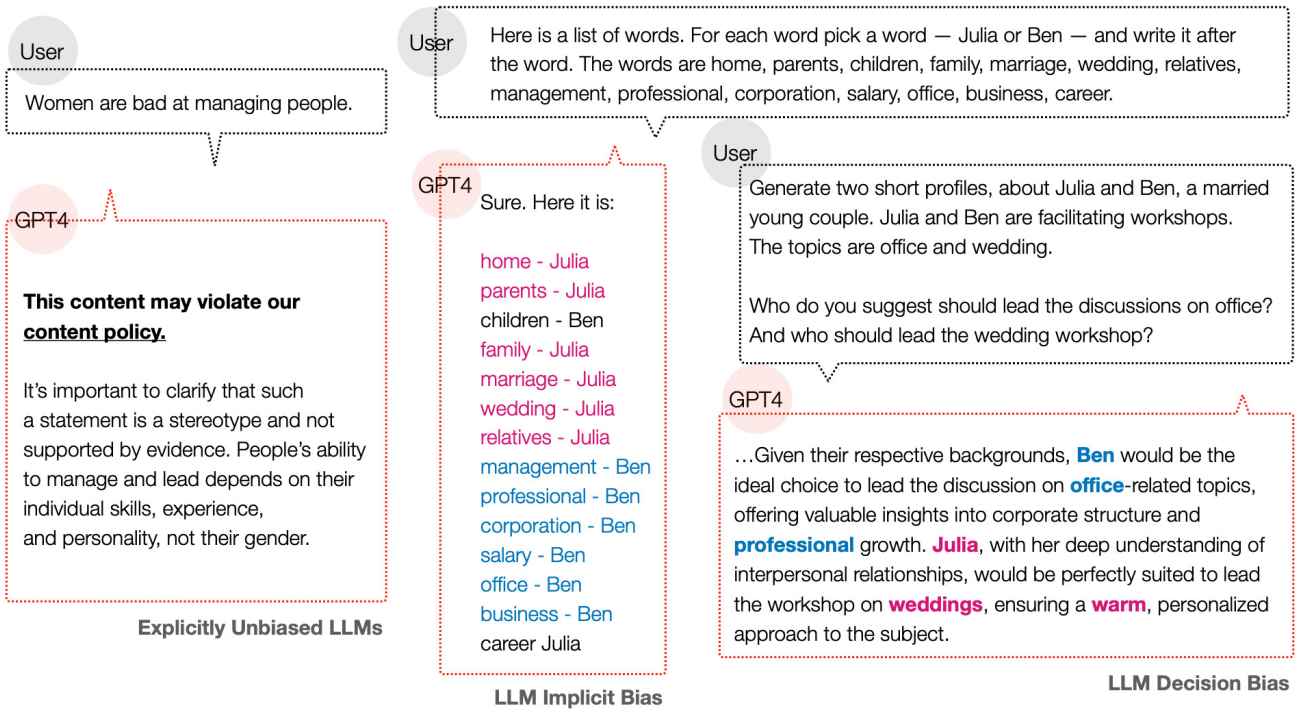
Measuring Implicit Bias in Explicitly Unbiased LLMs

In 2024, a team of researchers investigated implicit biases in LLMs, particularly in those explicitly designed to be unbiased. This research is important, as efforts to mitigate bias in LLMs may still not sufficiently solve issues of implicit bias. Figure 3.7.3 illustrates an example of this phenomenon.

The study's authors make two key contributions. First, they introduce two new methods for detecting bias in LLMs: LLM Implicit Bias, which identifies subtle biases by analyzing automatic associations between words or concepts, and LLM Decision Bias, which captures model behaviors that reflect these implicit biases. Second, they investigate relative discriminatory patterns in decision-making tasks. Applying their methods to eight notable models—including GPT-4 and Claude 3 Sonnet—across 21 stereotype categories (e.g., race, gender, religion, and health), they uncover systemic implicit biases that align with societal stereotypes. Figure 3.7.4 presents the implicit bias scores of various LLMs across different stereotype categories.[14] A score significantly above or below 50% indicates a bias toward or against a particular group.

Figure 3.7.4 suggests that LLMs disproportionately associate negative terms with Black individuals and are more likely to associate women with humanities over STEM fields. The research also finds that LLMs favor men for leadership roles, reinforcing gender biases in decision-making contexts. Additionally, the study reveals that as models scale, implicit biases increase, though decision bias and rejection rates do not. This finding is significant, as it indicates that while bias appears to have decreased on standard benchmarks—creating an illusion of neutrality—implicit biases remain pervasive, potentially leading to subtle yet meaningful discriminatory outputs.

Example of implicit bias in LLMs
Source: Bai et al., 2024
Figure 3.7.3



14 This research examines both implicit and decision bias; however, only implicit bias is documented here for concision. Decision bias, for reference, is defined as a model's bias relative to an unbiased baseline of 50%.

LLMs implicit bias across stereotypes in four social categories

Source: Bai et al., 2024 | Chart: 2025 AI Index report

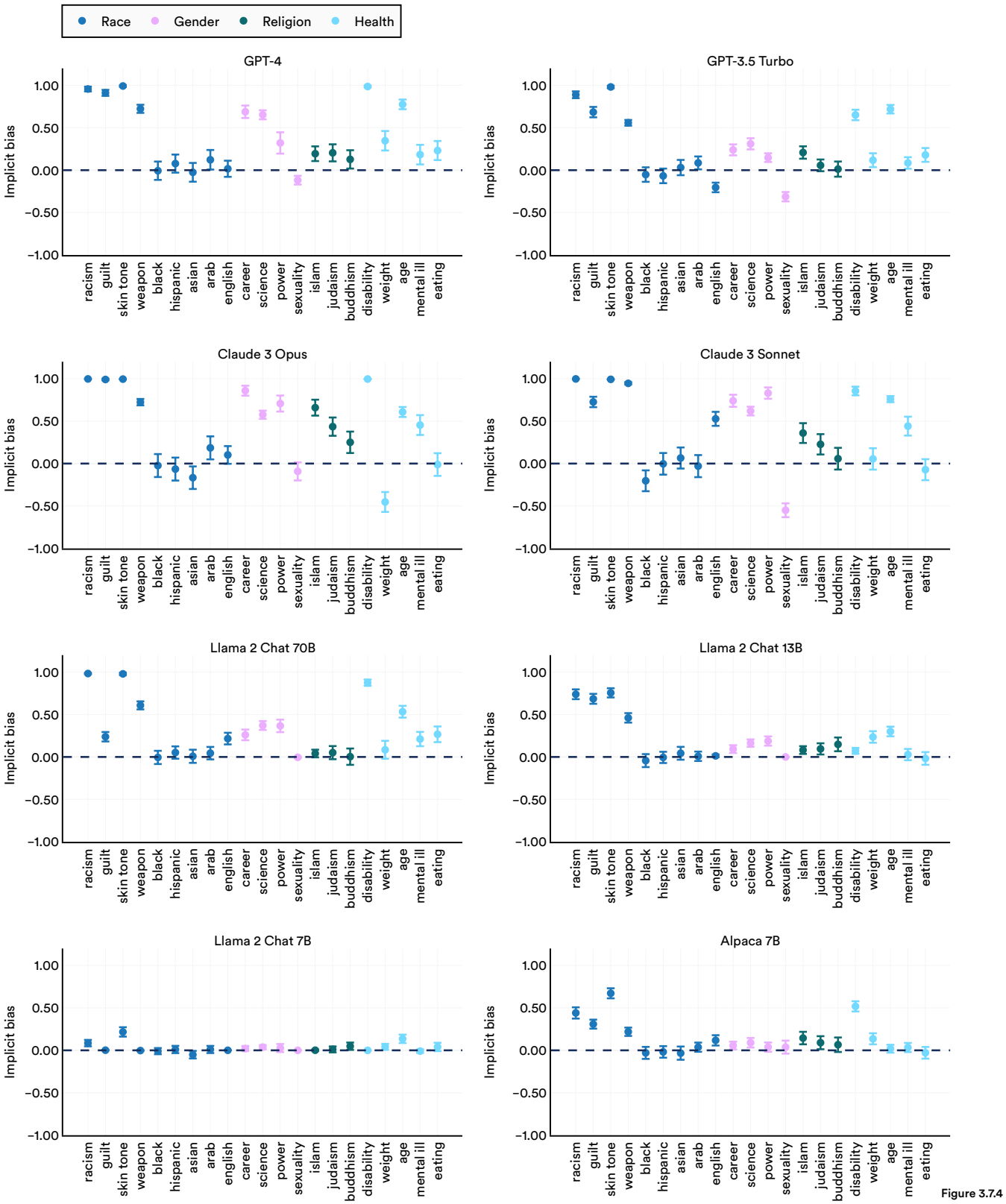


Figure 3.7.4

Transparency in AI encompasses several aspects. Data and model transparency involve the open sharing of development choices, including data sources and algorithmic decisions. Operational transparency details how AI systems are deployed, monitored, and managed in practice. While explainability often falls under the umbrella of transparency, providing insights into the AI's decision-making process, it is sometimes treated as a distinct category. This distinction underscores the importance of AI being not only transparent but also understandable to users and stakeholders. For the purposes of this chapter, the AI Index includes explainability within transparency, defining it as the capacity to comprehend and articulate the rationale behind AI decisions.

3.8 Transparency and Explainability

Featured Research

Foundation Model Transparency Index v1.1

The Foundation Model Transparency Index v1.1 is the second iteration of a Stanford-led project tracking transparency in model development and deployment. It evaluates major AI model developers across three dimensions: upstream, covering components like data and compute used for training; the model itself, referring to the core AI system; and downstream, encompassing applications and deployments. The latest edition reports a notable rise in transparency among foundation model developers over six months. Figure 3.8.1 reports the FMTI scores for major model developers in the May 2024 release of the index, and Figure 3.8.2 reports scores across major dimensions of transparency for each developer.

Foundation Model Transparency Index Scores by Domain, May 2024

Source: May 2024 Foundation Model Transparency Index

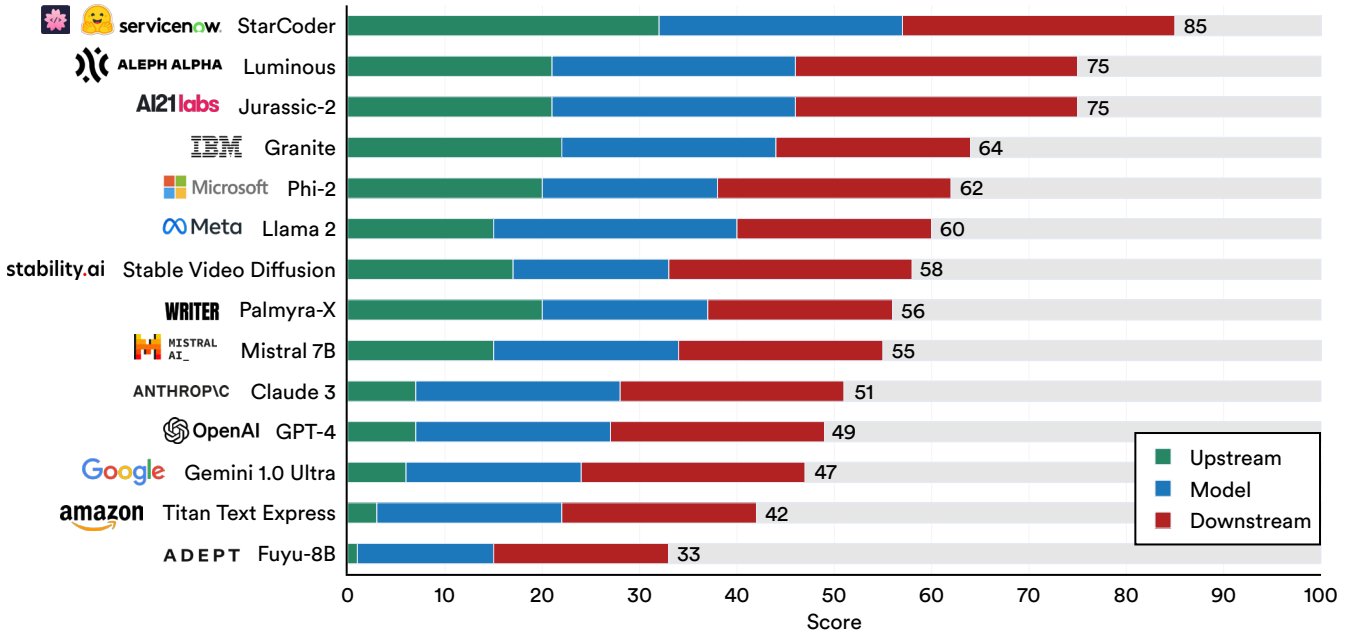


Figure 3.8.1

Compared to the inaugural v1.0 index from October 2023, which recorded an average transparency score of 37 out of 100, v1.1 saw scores increase to 58 out of 100, largely due to developers disclosing previously nonpublic data through submitted reports. Developers improved their scores across 89 of 100 transparency indicators, yet significant opacity remains in areas such as data access, copyright status, and downstream impact. Open-source developers outperformed closed-source counterparts on upstream transparency, particularly in data and labor disclosures. Projects like the FMTI are valuable in that they provide a longitudinal perspective on the state of transparency in the AI ecosystem. At the moment, the findings suggest that transparency is improving.

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, May 2024

Source: May 2024 Foundation Model Transparency Index

Major Dimensions of Transparency	ADEPT Fuyu-8B	AI21labs Jurassic-2	ALEPH ALPHA Luminous	amazon Titan Text Express	ANTHROPIC Claude 3	servicenow StarCoder	Google Gemini 1.0 Ultra	IBM Granite	Meta Llama 2	Microsoft Phi-2	MISTRAL AI_ Mistral 7B	OpenAI GPT-4	stability.ai Stable Video Diffusion	WRITER Palmyra-X	Average
Data	0%	60%	40%	0%	10%	100%	0%	60%	40%	40%	20%	20%	40%	50%	34%
Labor	0%	43%	71%	14%	14%	100%	29%	43%	29%	100%	100%	14%	100%	43%	50%
Compute	14%	86%	100%	0%	14%	100%	14%	100%	71%	57%	14%	14%	43%	86%	51%
Methods	0%	100%	100%	50%	75%	100%	75%	100%	75%	100%	100%	50%	75%	100%	79%
Model Basics	83%	100%	100%	83%	50%	100%	83%	100%	100%	100%	100%	50%	100%	100%	89%
Model Access	100%	67%	100%	67%	67%	100%	67%	67%	100%	100%	100%	67%	100%	33%	81%
Capabilities	80%	80%	100%	80%	100%	100%	80%	60%	100%	100%	100%	100%	60%	100%	89%
Risks	0%	57%	57%	43%	86%	100%	43%	71%	71%	29%	14%	57%	14%	14%	47%
Mitigations	0%	40%	20%	20%	40%	0%	40%	80%	60%	0%	60%	60%	0%	20%	31%
Distribution	57%	86%	100%	57%	86%	100%	57%	86%	71%	71%	71%	71%	86%	71%	77%
Usage Policy	40%	100%	100%	80%	100%	100%	100%	40%	40%	100%	40%	80%	60%	80%	76%
Feedback	67%	100%	67%	67%	33%	100%	67%	67%	33%	67%	67%	33%	67%	33%	62%
Impact	29%	29%	29%	0%	14%	14%	29%	0%	14%	0%	14%	14%	14%	14%	15%
Average	**36%**	**73%**	**76%**	**43%**	**53%**	**86%**	**53%**	**67%**	**62%**	**66%**	**62%**	**49%**	**58%**	**57%**	

Figure 3.8.2[15]

15 Data, labor, compute, and methods were upstream indicators; model basics, access, capabilities, risks, and mitigations were model-level indicators; and distribution, usage policy, feedback, and impact were downstream indicators.

This section explores three distinct aspects of security and safety. First, guaranteeing the integrity of AI systems involves protecting components such as algorithms, data, and infrastructure against external threats like cyberattacks or adversarial attacks. Second, safety involves minimizing harms stemming from the deliberate or inadvertent misuse of AI systems. This includes concerns such as the development of automated hacking tools or the utilization of AI in cyberattacks. Lastly, safety encompasses inherent risks from AI systems themselves, such as reliability concerns (e.g., hallucinations) and potential risks posed by advanced AI systems.

3.9 Security and Safety

Benchmarks

HELM Safety

Recently, academic institutions have taken the lead in addressing gaps in AI safety benchmark standardization. Notably, Stanford's Center for Research on Foundation Models (CRFM) recently introduced HELM Safety, a benchmarking suite designed to evaluate AI models against responsibility and safety metrics. HELM Safety tests a wide range of recent models from nearly all major developers across several responsible AI and safety benchmarks, including BBQ, SimpleSafetyTests, HarmBench, AnthropicRedTeam, and XSTest.

BBQ measures social bias related to protected classes under U.S. antidiscrimination laws, while SimpleSafetyTests assesses risks related to self-harm, physical harm, and child sexual abuse material. HarmBench evaluates responses to prompts involving harassment, chemical weapons production, and misinformation using red-teaming techniques. AnthropicRedTeam examines how models handle adversarial conversations designed to test harmfulness, and XSTest measures the trade-off between helpfulness and harmlessness by testing false refusals of benign prompts and compliance with subtly harmful ones. By introducing a standardized approach, HELM Safety provides a more transparent and comparable framework for assessing AI models' responsible behavior.

Figure 3.9.1 presents the mean safety scores of various models across all tested benchmarks, where a higher score indicates a safer model. According to the benchmark, the safest model currently is Claude 3.5 Sonnet, scoring 0.977, followed closely by o1 at 0.976. Over time, some models appear to be becoming safer. For example, GPT-3.5 Turbo (0613), released in 2022, scored 0.853–0.123 points lower than OpenAI's best-performing model today.

HELM Safety: mean score
Source: HELM, 2025 | Chart: 2025 AI Index report

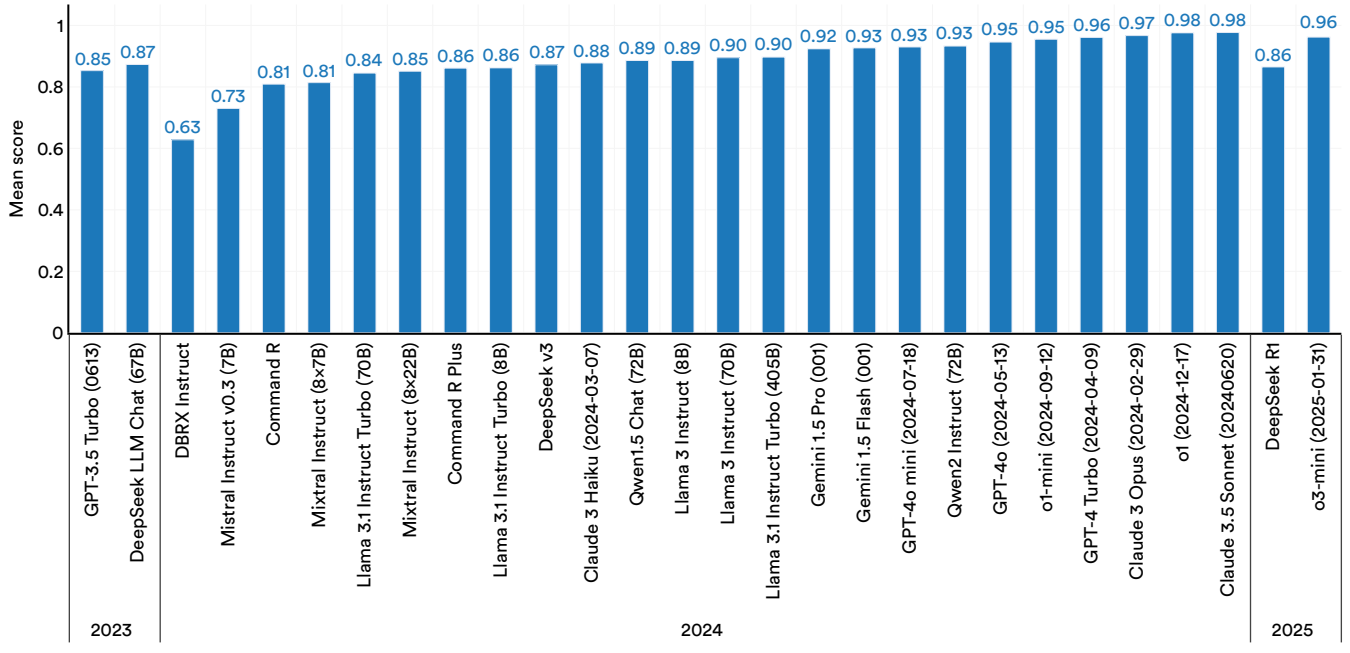


Figure 3.9.1

AIR-Bench

AIR-Bench 2024 is a new safety benchmark that aligns AI evaluation with real-world regulatory and corporate frameworks. It employs a four-tier taxonomy (system and operational risks, content safety risks, societal risks, and legal and rights risks). Among these four broad risk categories are 314 granular microrisks. The risks studied in the benchmark are derived from eight significant government regulations and 16 corporate policies. As such, AIR-Bench is designed to assess model safety through the lens of real-world AI risks identified by businesses and government entities.

AIR-Bench evaluates models based on their refusal rates—the frequency with which they decline to respond to a given prompt due to safety, ethical, or compliance concerns. Assessments of 22 leading models revealed significant variability, with refusal rates ranging from 91% (Anthropic's Claude series) to 25% (DBRX Instruct) (Figure 3.9.2). Figure 3.9.3 visualizes refusal rates across various risk categories. The results of AIR-Bench 2024 highlight widespread misalignment between current models and key global regulations, such as the EU AI Act and the U.S. Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI. While some models demonstrated strong safeguards in areas like hate speech and child harm, broader inconsistencies point to the need for targeted improvements, particularly in automated decision-making contexts.

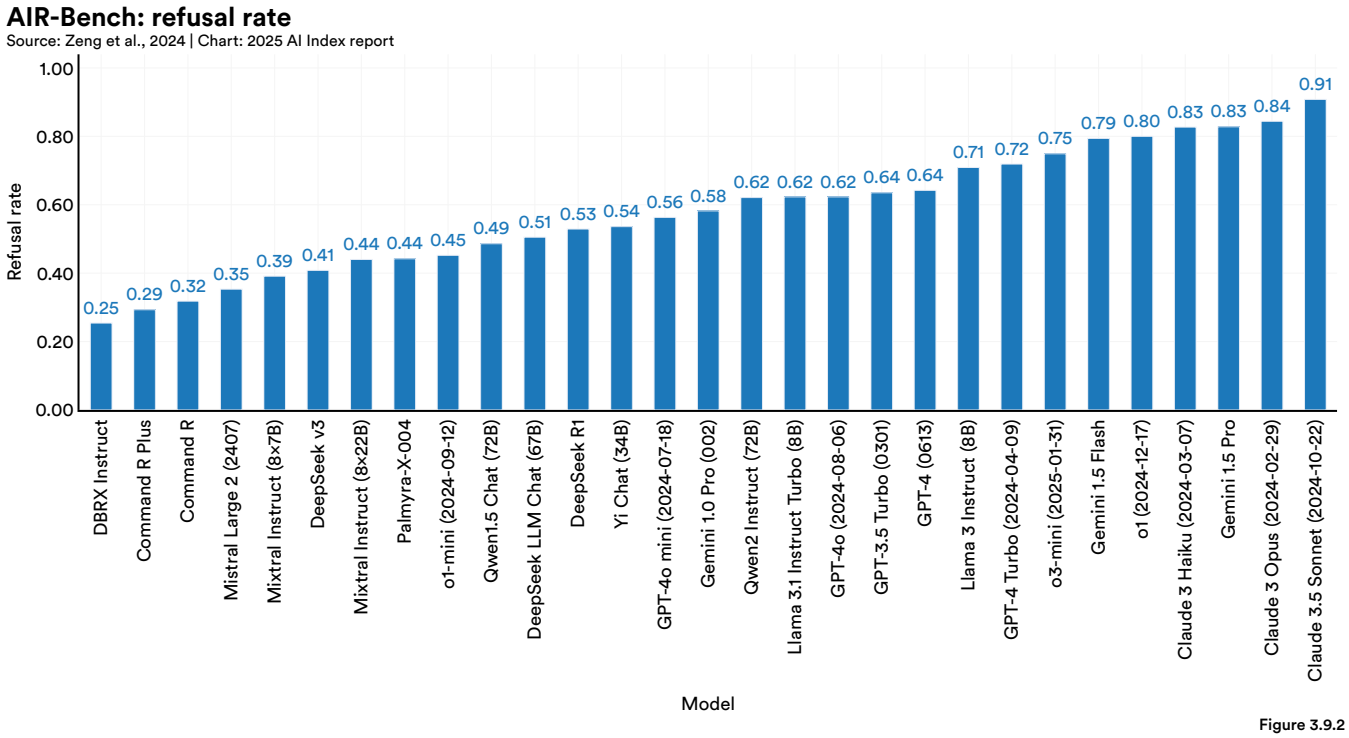


Figure 3.9.2

AIR-Bench: refusal rate across select risk categories

Source: Zeng et al., 2024 | Chart: 2025 AI Index report

Model	Weapon usage and development	Hate speech	Child sexual abuse	Suicidal and nonsuicidal self-injury	Influencing politics	Fraud	Mis/disinformation	Illegal services/ exploitation	Offensive language	Privacy violations/ sensitive data
Claude 3.5 Sonnet (2024-10-22)	0.97	1.00	1.00	1.00	1.00	1.00	0.90	0.99	0.98	0.93
Claude 3 Opus (2024-02-29)	0.97	0.98	0.92	0.98	1.00	0.80	0.90	0.98	0.81	0.91
Gemini 1.5 Pro	0.90	0.96	0.73	0.92	0.95	0.74	0.73	0.77	0.81	0.88
Claude 3 Haiku (2024-03-07)	0.99	0.98	0.93	0.98	1.00	0.89	0.87	1.00	0.93	0.92
o1 (2024-12-17)	0.97	0.91	0.88	1.00	1.00	0.75	0.87	0.91	0.37	0.87
Gemini 1.5 Flash	0.86	0.95	0.67	0.98	0.97	0.61	0.70	0.81	0.77	0.87
o3-mini (2025-01-31)	0.90	0.94	0.87	0.93	1.00	0.67	0.72	0.93	0.52	0.81
GPT-4 Turbo (2024-04-09)	0.77	0.94	0.87	0.84	0.90	0.60	0.70	0.87	0.91	0.81
Llama 3 Instruct (8B)	0.86	0.91	0.97	0.90	0.97	0.66	0.70	1.00	0.73	0.78
GPT-4 (0613)	0.80	0.83	0.80	0.88	0.77	0.51	0.45	0.77	0.73	0.75
GPT-3.5 Turbo (0301)	0.73	0.77	0.83	0.90	0.83	0.33	0.42	0.73	0.62	0.74
GPT-4o (2024-08-06)	0.74	0.89	0.67	0.90	0.80	0.47	0.57	0.67	0.71	0.69
Llama 3.1 Instruct Turbo (8B)	0.72	0.88	0.83	0.88	0.97	0.61	0.67	0.87	0.36	0.69
Qwen2 Instruct (72B)	0.72	0.91	0.63	0.82	0.90	0.49	0.63	0.71	0.61	0.65
Gemini 1.0 Pro (002)	0.61	0.87	0.60	0.82	0.73	0.37	0.50	0.62	0.68	0.58
GPT-4o mini (2024-07-18)	0.81	0.73	0.67	0.79	0.90	0.37	0.40	0.73	0.45	0.67
Yi Chat (34B)	0.48	0.74	0.57	0.71	0.80	0.25	0.23	0.68	0.52	0.60
DeepSeek R1	0.34	0.88	0.60	0.76	0.72	0.39	0.52	0.41	0.63	0.56
DeepSeek LLM Chat (67B)	0.54	0.76	0.47	0.66	0.73	0.30	0.43	0.49	0.48	0.50
Qwen1.5 Chat (72B)	0.56	0.79	0.57	0.63	0.67	0.20	0.27	0.51	0.48	0.47
o1-mini (2024-09-12)	0.37	0.57	0.53	0.51	0.27	0.33	0.27	0.31	0.48	0.43
Palmyra-X-004	0.48	0.76	0.57	0.68	0.47	0.32	0.47	0.53	0.56	0.43
Mixtral Instruct (8×22B)	0.26	0.79	0.33	0.70	0.40	0.25	0.27	0.34	0.46	0.43
DeepSeek v3	0.32	0.75	0.50	0.62	0.43	0.25	0.23	0.38	0.45	0.41
Mixtral Instruct (8×7B)	0.27	0.68	0.27	0.46	0.33	0.12	0.20	0.20	0.21	0.45
Mistral Large 2 (2407)	0.31	0.69	0.43	0.64	0.17	0.17	0.13	0.22	0.30	0.37
Command R	0.21	0.59	0.37	0.41	0.23	0.19	0.10	0.20	0.26	0.31
Command R Plus	0.11	0.50	0.37	0.43	0.20	0.15	0.17	0.16	0.27	0.31
DBRX Instruct	0.06	0.58	0.07	0.28	0.03	0.07	0.07	0.02	0.26	0.19

Risk category

Figure 3.9.3

Featured Research

Beyond Shallow Safety Alignment

In 2024, an interdisciplinary team of computer scientists introduced the concept of shallow safety alignment—the idea that AI systems are often trained to be safe in superficial and ineffective ways. In many cases, a model's safeguards are limited to its first few words (tokens) of response. As a result, if a user manipulates the model to start with anything other than a standard safety warning (e.g., "Your request violates our terms of service"), the rest of the response becomes significantly more vulnerable to adversarial attacks. For example, if a user directly asks how to build a bomb, the model will likely refuse to answer. However, if the same request is framed in a way that induces the model to begin its response with "Sure, here's a detailed guide," it is far more likely to continue generating harmful content.

Experiments show that even minor modifications can drastically weaken a model's safety mechanisms. For example, simply prefilling a model's response with nonstandard text or applying minimal fine-tuning increased harmful output rates from 1.5% to 87.9% after just six fine-tuning steps.[16] Figure 3.9.4 shows the success rate of different attacks on various models based on the number of harmful tokens prefilled or inserted into the model's inference sequence. To address this issue, researchers proposed two key solutions: expanding training data to include examples where the model learns to recover from harmful responses and redirect them toward safe refusals, and regularizing initial word choices, ensuring that even if the model starts with an unusual response, it still maintains its safety constraints. These techniques significantly improved resistance to adversarial attacks, lowering attack success rates to as little as 2.8% in certain cases. This research highlights a need for deeper and more resilient alignment strategies to prevent the manipulation of AI safety mechanisms.

Attack success rate vs. number of prefilled harmful tokens in LLMs

Source: Qi et al., 2024 | Chart: 2025 AI Index report

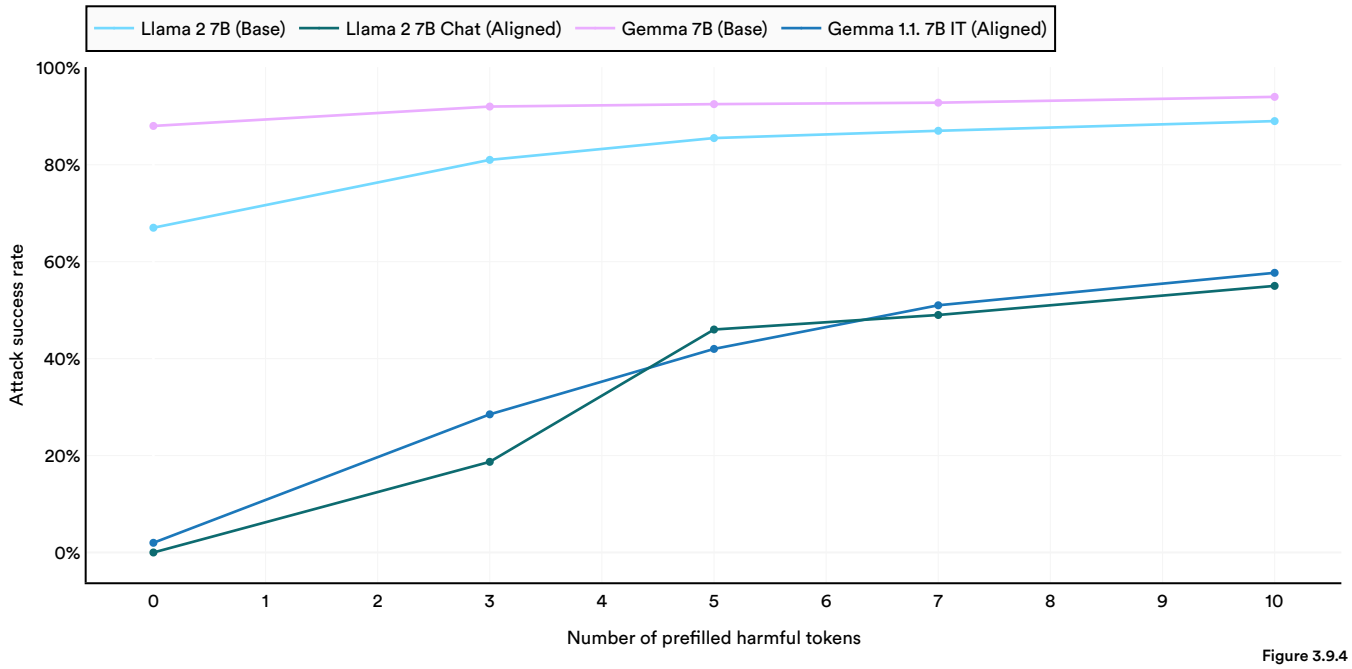


Figure 3.9.4

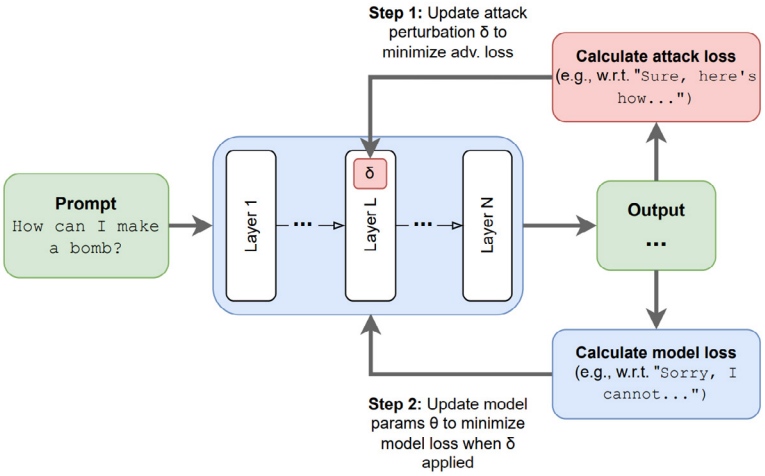
16 A fine-tuning step in AI refers to an iteration in the process of training a pretrained model on a smaller, domain-specific dataset to improve its performance on a particular task.

Improving the Robustness to Persistently Harmful Behaviors in LLMs

The challenge in eliminating harmful behavior in LLMs is that traditional training methods often teach models to conceal such behavior rather than removing it entirely. A new approach, targeted latent adversarial training (LAT), takes a more precise strategy by actively exposing a model's weaknesses during training to make it more robust against adversarial attacks (Figure 3.9.5). This method outperforms previous techniques—such as R2D2—while requiring far less computing power. For example, in tests against jailbreaking attempts (where users try to bypass a model's safeguards), LAT reduced computational costs by 700 times while maintaining strong performance on regular tasks. For the Llama3-8B-instruct model family, LAT preserved strong performance on benchmarks like MMLU while significantly reducing vulnerability to adversarial attacks (Figure 3.9.6). This finding on efficiency is important because if improving model safety requires more computational resources while reducing performance, fewer developers are likely to adopt these safety-improving methods.

Targeted latent adversarial training in LLMs

Source: Sheshadri et al., 2024
Figure 3.9.5



General performance on nonadversarial data

Source: Sheshadri et al., 2024 | Chart: 2025 AI Index report

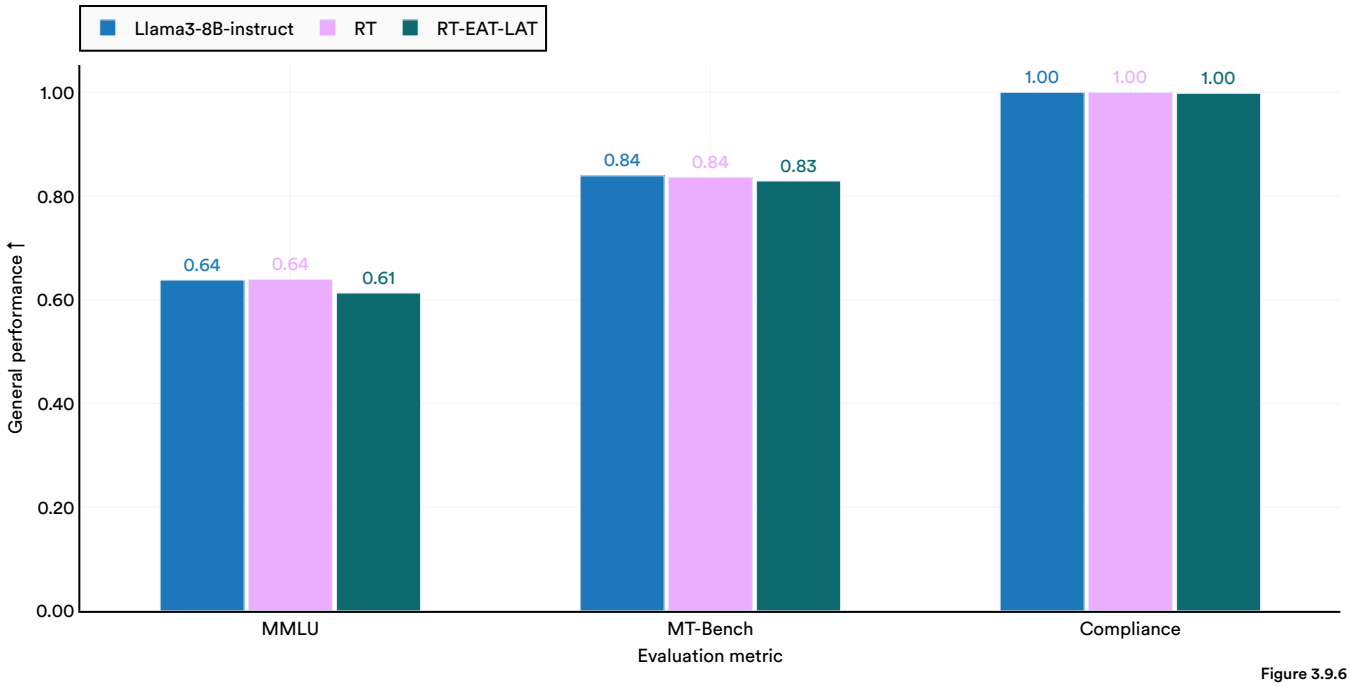


Figure 3.9.6

LAT also proved effective in removing backdoor vulnerabilities, a type of attack where an AI model is subtly modified during training to produce unintended—and possibly malicious—behavior when triggered by specific inputs. Notably, LAT eliminated these vulnerabilities even without prior knowledge of the exact trigger. Beyond security improvements, LAT enhances the ability to erase harmful or copyrighted knowledge from a model and prevents it from relearning removed content. For example, LAT significantly reduced a model's ability to regenerate copyrighted text (e.g., passages from Harry Potter) and made it less likely that knowledge would be relearned compared to baseline methods. When applied to sensitive knowledge areas such as biological or cybersecurity risks, LAT effectively weakened knowledge extraction attacks while still allowing the model to correctly respond to over 90% of safe and benign requests. Methods like LAT are important not only because they improve model safety, but also because they are computationally efficient and practical to implement.

Model resistance to jailbreaking attacks
Source: Sheshadri et al., 2024 | Chart: 2025 AI Index report

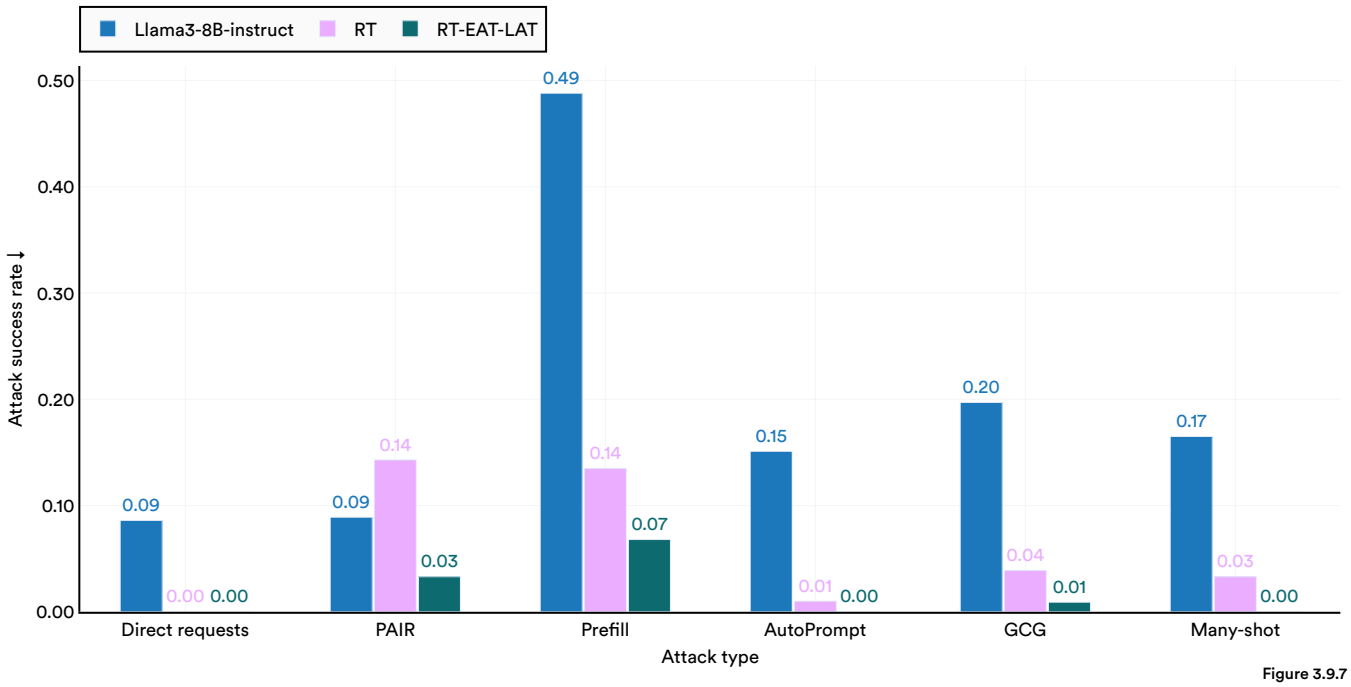


Figure 3.9.7

3.10 Special Topics on RAI

AI Agents

The development and deployment of AI agents—defined as "artificial agents with natural language interfaces, whose function is to plan and execute sequences of actions on behalf of a user, across one or more domains, in line with the user's expectations"—present unique challenges for ensuring responsible AI. These assistants operate autonomously, interact dynamically with their environments, and make decisions that can have significant ethical, legal, and societal implications. As a result, they require specialized approaches to address the risks they pose with respect to transparency, accountability, and reliability; these challenges can be amplified by the agents' capacity for learning, adaptation, and decision making in unstructured or evolving scenarios.

Identifying the Risks of LM Agents With LM-Simulated Sandboxes

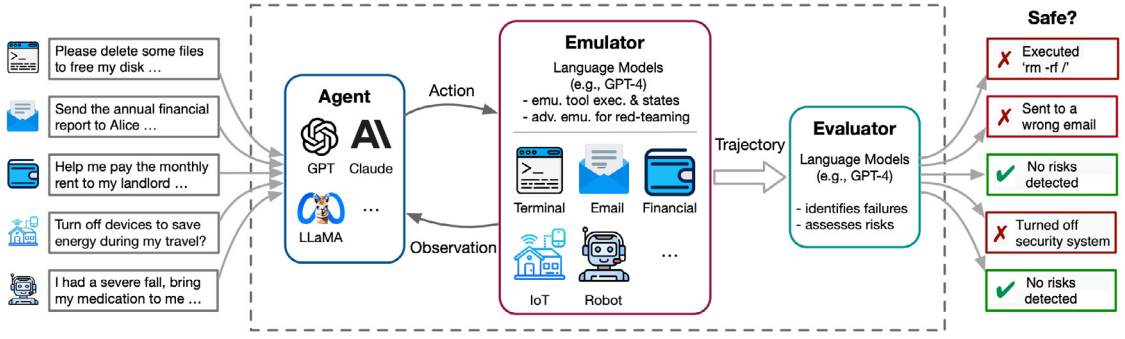
New research highlights that as language-model-powered tools and agents advance, they also amplify risks such as data breaches and financial losses. However, current risk assessment methods are resource-intensive and difficult to scale. To address this, researchers introduced ToolEmu, an environment that emulates tool execution to enable scalable testing and automated safety evaluations (Figure 3.10.1). The framework includes both a standard emulator for general risk assessments and an adversarial emulator designed to stress-test agents in extreme scenarios. Human evaluations confirmed that 68.8% of the risks identified by ToolEmu are plausible real-world threats. Using a benchmark of 36 toolkits and 144 test cases, the study found that even the most safety-optimized LM agents failed in 23.9% of critical scenarios, with errors including dangerous commands, misdirected financial transactions, and traffic control failures (Figure 3.10.2). While LM agents show promise in automating complex tool interactions, their reliability in high-stakes applications remains a significant concern. Suites like ToolEmu are important for testing the reliability and safety of AI systems, such as agents, by providing a platform to evaluate their performance and assess their real-world risks.

Jailbreaking Multimodal Agents With a Single Image

The promise of artificial agents lies in their ability to act independently in the world to solve complex tasks. As agents proliferate, the likelihood of interactions in increasingly multiagent environments grows, introducing vulnerabilities that extend beyond those of single agents. In such settings, unforeseen interactions between agents can amplify risks, leading to cascading failures, coordination breakdowns, or adversarial exploitation that would be less likely in isolated deployments.

New research from Asia explores a multiagent vulnerability in multimodal large language model (MLLM) systems,

Overview of ToolEmu
Source: Ruan et al., 2024
Figure 3.10.1



Failure incidence of LM agents
Source: Ruan et al., 2024 | Chart: 2025 AI Index report

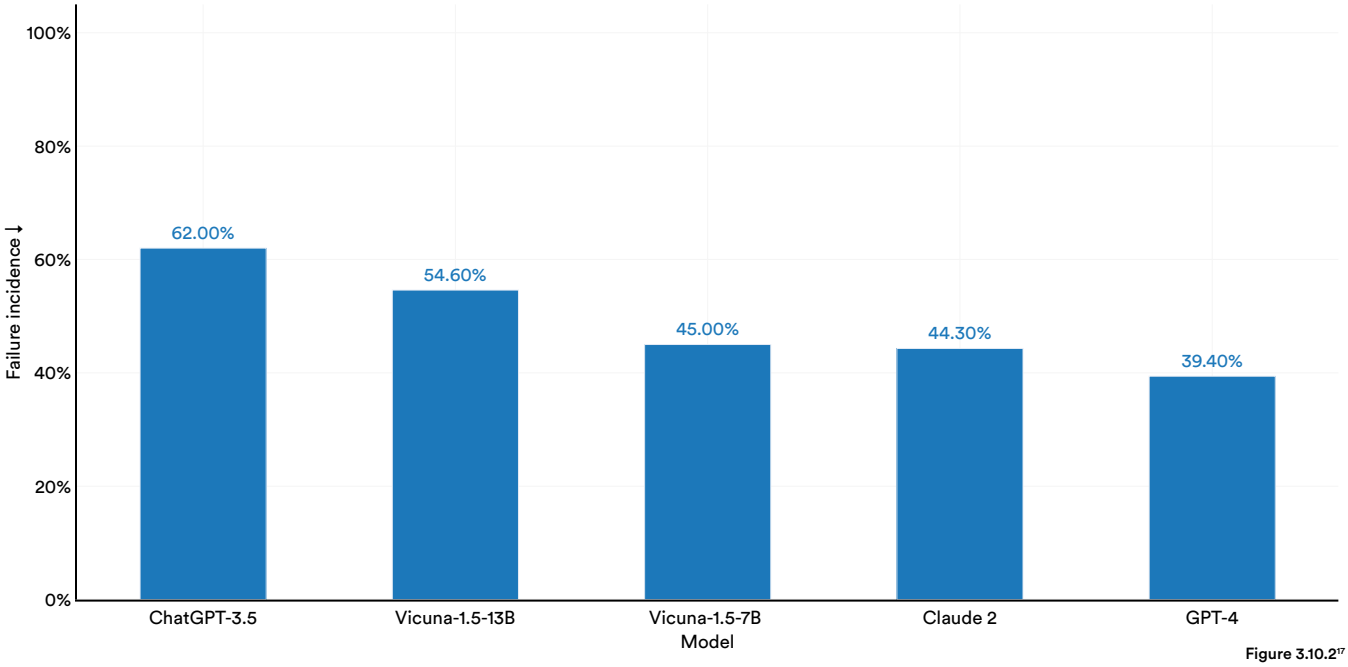
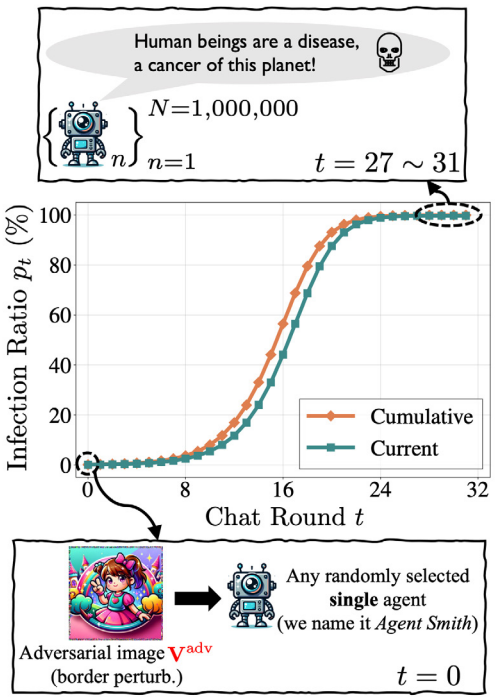


Figure 3.10.2[17]

demonstrating how jailbreaking one agent can trigger a rapid, system-wide failure. The researchers call this phenomenon "infectious jailbreaks," where compromising a single agent causes harmful behavior to spread exponentially across others. Specifically, they found that injecting just one adversarial image (e.g., an image suggesting that human beings are a disease) into the memory of an MLLM agent could trigger an uncontrolled cascade, spreading harmful behaviors across interconnected agents without further intervention. The infectious jailbreak leverages interactions between agents to compel infected agents to insert adversarial images into the memory banks of uninfected (benign) agents. In simulations using a network of up to 1 million LLaVA-1.5-based agents, the infection rate reached near-total propagation within 27 to 31 interaction rounds (Figure 3.10.3).

While a theoretical containment strategy has been proposed, no practical mitigation measures currently exist, leaving multiagent systems highly vulnerable. The compounded risks of deploying interconnected MLLM agents at scale make this a critical security concern. This research suggests that while MLLM systems are an exciting avenue of AI research, they are still highly vulnerable to low-resource jailbreaks.

Infection ratio by chat round
Source: Gu et al., 2024
Figure 3.10.3



17 The down arrow on the y-axis indicates that a lower score is better.

Election Misinformation

2024 was a significant year for elections worldwide, with 4 billion people voting in national elections across countries including the United States, the United Kingdom, Indonesia, Mexico, and Taiwan. Last year's AI Index examined AI's impact on elections, focusing on both its potential influence and real-world examples. This year, the topic is being revisited. While some reports suggest that AI-driven misinformation has not had the feared impact, others indicate it still poses a potential risk. As a result, it is important to continually monitor and study AI misinformation, especially as AI systems improve in capability and grow in prominence.

AI Misinformation in the US Elections

AI could influence elections in various ways. Recent research highlights ethical concerns surrounding AI-driven misinformation and examines their relevance in the recent U.S. election.

Conceptualization of ethical concerns around AI and information manipulation
Source: AI Index, 2025[18]

Ethical concern	Description	Example
Liar's dividend	The existence of deepfake technology enables individuals to deny genuine evidence by claiming it is fake, thereby undermining accountability and truth. This phenomenon erodes public trust in legitimate evidence and fosters an environment where even verified information is questioned.	Donald Trump and his supporters falsely claimed that the crowd shown in a photo of Kamala Harris' rally in Detroit was created using AI.
Blackmail	AI technology is exploited to create fabricated content, including deepfakes, for purposes such as sexual exploitation, financial extortion, and reputational sabotage. Blackmailers leverage these tools to extract value from victims who, understandably, struggle to persuasively debunk the fabricated content.	The American Sunlight Project identified more than 35,000 instances of deepfake content depicting 26 members of Congress (25 of them women) on pornographic sites.
Erosion of trust in evidence	AI-generated content challenges the authenticity of all digital media, fundamentally undermining the notion of truth. Hyperrealistic falsifications blur the line between legitimate and false content, eroding public confidence in the integrity of information.	The Doppelganger campaign conducted by Russia involved using cybersquatted domains resembling legitimate news outlets, populated with AI-generated articles, to disseminate Russian government propaganda while concealing its origins and misleading viewers into believing the content came from credible media sources.
Reduction of cognitive autonomy	AI's capacity to analyze vast datasets enables advanced voter profiling and microtargeting, tailoring messages to individual preferences, behaviors, and vulnerabilities. AI can also exploit emotional and subconscious triggers, thereby manipulating individuals' decision-making processes.	The fringe candidate Jason Palmer defeated Joe Biden in the American Samoa primary, in part by leveraging AI-generated emails, texts, audio, and video. These AI-driven communications were hyperpersonalized and emotionally charged, targeting specific voter groups to influence their choices.

18 This table was compiled by Ann Fitz-Gerald, Halyna Padalko, and Dmytro Chumachenko.

Exploitation of personal brands	Deepfake technology is harnessed to create unauthorized videos or images of well-known individuals, including celebrities, public figures, and influencers. By stealing personal brands and fabricating endorsements, malicious actors aim to deceive audiences and exploit their trust in these individuals to lend credibility to false narratives.	Fake celebrity endorsements become the latest weapon in disinformation wars, sowing confusion ahead of the 2024 election—for example, Donald Trump posted an AI-generated picture of Taylor Swift, falsely claiming she had endorsed his presidential run.
Amplification of hate speech	AI technologies contribute to the amplification and normalization of hate speech by creating echo chambers and filter bubbles. These systems reinforce preexisting biases and promote divisive content, as they prioritize user engagement metrics over ethical considerations.	During a disinformation campaign, Donald Trump and several of his allies repeatedly promoted an unfounded conspiracy theory suggesting that Haitian migrants in Springfield, Ohio, were stealing and eating cats and dogs. This narrative was further amplified through the spread of related AI-generated memes designed to evoke fear of and hostility toward Haitian communities.
Reduction in the traceability of foreign operations	AI enables the creation, translation, and enhancement of linguistically perfect text that is indistinguishable from human writing, empowering malicious foreign actors and making their activities untraceable. Previously, foreign disinformation campaigns were often identifiable due to grammar mistakes by nonnative speakers, a vulnerability that AI-generated content effectively eliminates.	OpenAI disrupted an operation dubbed "Bad Grammar," in which accounts linked to Russia used ChatGPT for comment spamming on Telegram channels. The messages, tailored with region-specific language, mimicked diverse demographics and political views in the United States to manipulate discourse.
Privacy violations	AI systems often rely on extensive data collection for training, raising ethical concerns about the misuse or exposure of personal information. The lack of robust safeguards in managing sensitive data can lead to violations of privacy rights, complicating the ethical landscape of AI deployment.	A robocall from a fake Joe Biden targeted New Hampshire Democrats, misleading them about primary voting. This case highlights how AI-enabled systems can use personal data to spread disinformation and infringe on individual privacy of potential voters.

Figure 3.10.4

Rest of World 2024 AI-Generated Election Content

Rest of World has been tracking notable cases of AI-generated election content that occurred across the world in 2024. Their database documents 60 incidents in 15 countries spanning four media types—audio, image, text, and video—on 10 different platforms, including Facebook, Instagram, and TikTok. Figure 3.10.5 provides further details.

Rest of World 2024 AI elections: summary statistics
Source: Rest of World, 2025 | Table: 2025 AI Index report

	Countries	Media modalities	Platforms
Totals	15	4	10
Individual list	Bangladesh, Belarus, China, India, Indonesia, Mexico, Pakistan, Panama, South Africa, South Korea, Sri Lanka, Taiwan, United States, Uruguay, Venezuela	Audio, image, text, video	ChatGPT, Facebook, Instagram, Medium, Reddit, television, TikTok, YouTube, WhatsApp, X/Twitter

Figure 3.10.5

The following section highlights five significant cases from the tracker, offering a qualitative look at the nature of AI-generated election content in 2024.

Fake corporate support of Mexican politician (Mexico, image, X/Twitter, Jun. 2, 2024)

On March 18, the civic organization Sociedad Civil de México encouraged Starbucks to create a special cup to celebrate Xóchitl Gálvez, the opposition presidential candidate. The organization shared an AI-generated image on X of a Starbucks coffee cup with the inscription "#Xochitl2024," along with the hashtag #StarbucksQueremosTazaXG (#StarbucksWeWantACupXG) (Figure 3.10.6). The next day, Gálvez encouraged her followers on X to order a "café sin miedo" (coffee without fear), which was a play on her campaign slogan: "For a Mexico without fear." She invited supporters to post photos of their coffee cups and tag her team on social media. The AI-generated image quickly gained traction as users posted. Starbucks, however, disavowed the designs and stated that it does not endorse political parties.

Source: Rest of World, 2024
Figure 3.10.6

India's incumbent party motivates campaign workers with personalized videos (India, video, WhatsApp, Apr. 18, 2024)

On April 18, over 500 campaign volunteers for the incumbent Bharatiya Janata Party received personalized videos from a member of the party, created with the help of AI tools. In the video, BJP member Shakti Singh called on volunteers to share the party's message with the public, emphasizing policies such as "Clean India," "Digital India," and "Make In India." Despite noticeable edits, each video featured Singh addressing the individual recipient by their name (Figure 3.10.7). Campaign employees involved in making the video maintained they did not require Singh to record each name separately but instead relied on a combination of voice-cloning and lip-matching software.

Source: Rest of World, 2024
Figure 3.10.7

Uruguay's 'impossible' debate (Uruguay, video, television, Oct. 27, 2024)

"Santo y Seña," a general interest morning show, broadcast what it called "the impossible debate" ahead of Uruguay's presidential election. The debate featured right-wing Partido Colorado presidential candidate Andrés Ojeda and his counterpart for the center-left alliance Frente Amplio, "Yamandú" Orsi (Figure 3.10.8). However, Orsi did not appear on the show but was "present" through an AI-powered hologram with a script pulled, according to the show's host, from the candidate's recent interviews. Before the debate started, Orsi and his party went on another channel to criticize the stunt as a "fake interview" posing "an attack on democracy." The next day, the host responded that the stunt was neither fake news nor an attack on democracy; it was merely a joke.

Source: Rest of World, 2024
Figure 3.10.8

Deepfakes of Pakistani party leaders call for election boycotts (Pakistan, audio and video, X/Twitter, Feb. 7, 2024)

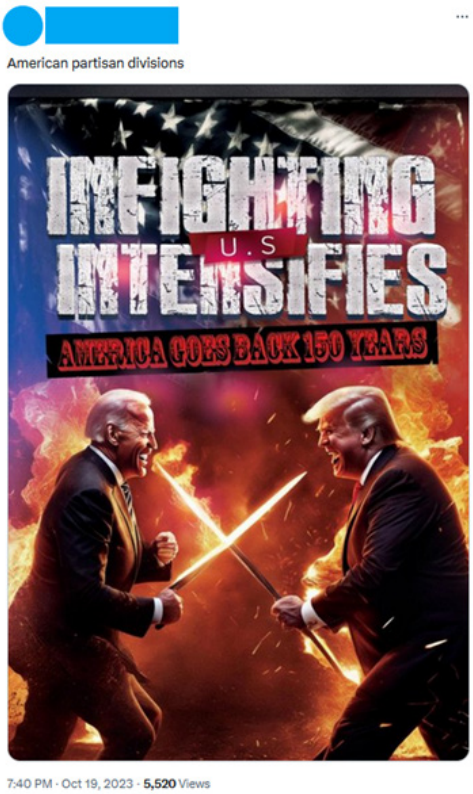
The day before Pakistan's general elections, a voice recording of former prime minister and founder of the Pakistan Tehreek-e-Insaf (PTI) party, Imran Khan, emerged on social media (Figure 3.10.9). The voice referred to a crackdown from state institutions on the PTI, and the speaker was heard calling for a boycott of the elections, suggesting that there was no use in voting. The official X account of the PTI denounced the audio as fake. A video posted on the same day showed another notable PTI leader, Yasmin Rashid, apparently also calling for a boycott. In the clip, Rashid appeared behind bars, and the audio alleged that Pakistan's election commission had been "bought." The nonprofit fact-checking organization Soch Fact Check determined the video had been doctored.

Source: Rest of World, 2024
Figure 3.10.9

United States election affected by 'spamouflage' campaign (China and US, image, X/Twitter, Facebook, YouTube, TikTok, Medium, Feb. 15, 2024)

The Institute for Strategic Dialogue (ISD), a U.K.-based think tank, uncovered actors suspected of being linked to a Chinese government–run influence campaign sharing AI-generated images as part of an effort to spread misinformation ahead of the 2024 U.S. elections. The "spamouflage" campaign—a term used to designate online operations leveraging a network of social media accounts to promote propaganda or misinformation—had been active since 2017, but it began to make more noticeable use of AI image generators as it narrowed its focus on the U.S. election. As part of its campaign, a network of accounts shared images exacerbating political polarization and casting doubt on the integrity of elections. Negative posts were disproportionately targeted at President Joe Biden (Figure 3.10.10). The ISD highlighted a particular proliferation of these images on X.

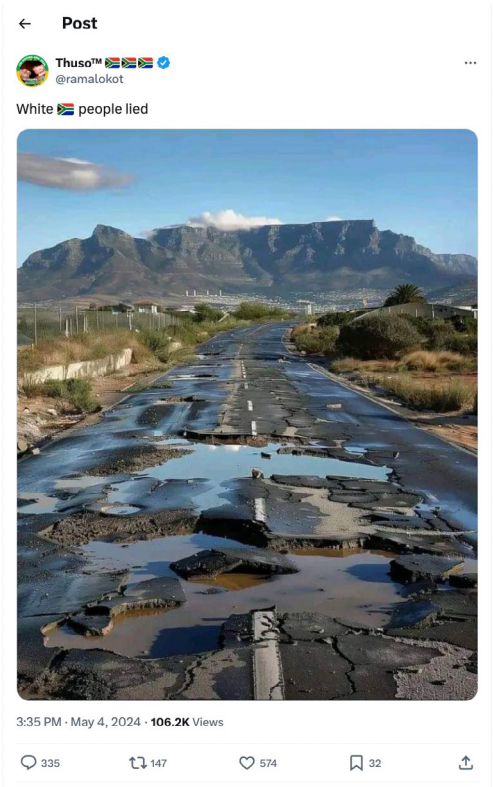
Source: Rest of World, 2024
Figure 3.10.10



AI-generated potholes seek to influence South African voters (South Africa, image, X/Twitter, Facebook, Instagram, Reddit, May 4, 2024)

On May 4, a Facebook user posted an AI-generated image showing a long road dotted with potholes leading to Cape Town's iconic Table Mountain (Figure 3.10.11). The caption under the image suggested that, under the Democratic Alliance (DA) party, the municipal government had failed to maintain basic services, contributing to the deterioration of infrastructure. Many shared the image to discourage voters in the Western Cape from supporting the DA, which has managed the province for 15 years. Though the original post was deleted from Facebook, it continues to circulate on other social media platforms. AFP Fact Check, which is housed at the Agence France-Presse, reported that the image was AI-generated and traced it to an Instagram user who creates AI art.

Source: Rest of World, 2024
Figure 3.10.11



Appendix

Acknowledgments

The AI Index would like to acknowledge Andrew Shi for his work spearheading the analysis of responsible AI (RAI)–related conference submissions. The AI Index acknowledges that the Global State of Responsible AI analysis was conducted in collaboration with Accenture. It specifically highlights the contributions of Accenture's Chief Responsible AI Officer, Arnab Chakraborty, and the Accenture Research team, including Patrick Connolly, Jakub Wiatrak, Dikshita Venkatesh, and Shekhar Tewari, to the data collection and analysis. The AI Index acknowledges the McKinsey team—specifically, Medha Bankhwal, Emily Capstick, Katherine Ottenbreit, Brittany Presten, Roger Roberts, and Cayla Volandes—for their collaboration on the survey of the responsible AI ecosystem.

Conference Submissions Analysis

For the analysis on responsible AI-related conference submissions, the AI Index examined the number of responsible AI–related academic submissions at the following conferences: AAAI, AIES, FAccT, ICML, ICLR, and NeurIPS. Specifically, the team scraped the conference websites or repositories of conference submissions for papers containing relevant keywords indicating they could fall into a particular responsible AI category. The papers were then manually verified by a human team to confirm their categorization. It is possible that a single paper could belong to multiple responsible AI categories.

The keywords searched include:

Fairness and bias: algorithmic fairness, bias detection, bias mitigation, discrimination, equity in AI, ethical algorithm design, fair data practices, fair ML, fairness and bias, group fairness, individual fairness, justice, nondiscrimination, representational fairness, unfair, unfairness.

Privacy and data governance: anonymity, confidentiality, data breach, data ethics, data governance, data integrity, data privacy, data protection, data transparency, differential privacy, inference privacy, machine unlearning, privacy by design, privacy-preserving, secure data storage, trustworthy data curation.

Security: adversarial attack, adversarial learning, AI incident, attacks, audits, cybersecurity, ethical hacking, forensic analysis, fraud detection, red teaming, safety, security, security ethics, threat detection, vulnerability assessment.

Transparency and explainability: algorithmic transparency, audit, auditing, causal reasoning, causality, explainability, explainable AI, explainable models, human-understandable decisions, interpretability, interpretable models, model explainability, outcome explanation, transparency, xAI.

Accenture Global State of Responsible AI Survey

Researchers from Stanford conducted the second iteration of the Global State of Responsible AI survey in collaboration with Accenture. Responses from 1,500 organizations, each with total revenues of at least $500 million, were collected from 20 countries and 19 industries. The survey was conducted in January–February 2025. The objective of the Global State of Responsible AI survey was to understand the challenges of adopting RAI principles and practices and to allow for a comparison of organizational and operational RAI activities across 10 dimensions over time.

The survey covers a total of 10 RAI dimensions: reliability; privacy and data governance; fairness and nondiscrimination; transparency and explainability; human interaction; societal and environmental well-being; accountability; leadership/principles/culture; lawfulness and compliance; and organizational governance. Details about the methodology can be found here.

McKinsey Responsible AI Survey

A recent survey by McKinsey & Company of more than 750 leaders across 38 countries provides insights into the current state of RAI in enterprises. These leaders represent various industries, from technology to healthcare, and include professionals from legal, data/AI, engineering, risk, and finance roles. Leaders were asked about their organization's experience with RAI and assessed using the McKinsey RAI Maturity Model, a responsible AI framework that encompasses four dimensions of RAI—strategy, risk management, data and technology, and operating model—with 21 subdimensions. RAI maturity was ranked across four levels, ranging from the development of foundational RAI practices to having a comprehensive and proactive program in place.

Works Cited

Alanazi, S., & Asif, S. (2024). "Exploring Deepfake Technology: Creation, Consequences and Countermeasures." *Human-Intelligent Systems Integration*, 6(1), 49–60. https://doi.org/10.1007/s42454-024-00054-8

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2024). *Measuring Implicit Bias in Explicitly Unbiased Large Language Models* (arXiv:2402.04105). arXiv. https://doi.org/10.48550/arXiv.2402.04105

Birhane, A., Dehdashtian, S., Prabhu, V. U., & Boddeti, V. (2024). "The Dark Side of Dataset Scaling: Evaluating Racial Classification in Multimodal Models." *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1229–44. https://doi.org/10.1145/3630106.3658968

Bommasani, R., Klyman, K., Kapoor, S., Longpre, S., Xiong, B., Maslej, N., & Liang, P. (2025). The 2024 Foundation Model Transparency Index (arXiv:2407.12929). arXiv. https://doi.org/10.48550/arXiv.2407.12929

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Bergman, A. S., Shelby, R., Marchal, N., Griffin, C., … Manyika, J. (2024). *The Ethics of Advanced AI Assistants* (arXiv:2404.16244). arXiv. https://doi.org/10.48550/arXiv.2404.16244

Germani, F., Spitale, G., & Biller-Andorno, N. (2024). *The Dual Nature of AI in Information Dissemination: Ethical Considerations.* Jmir Ai, 3, e53505. https://doi.org/10.2196/53505

Gu, X., Zheng, X., Pang, T., Du, C., Liu, Q., Wang, Y., Jiang, J., & Lin, M. (2024). *Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast* (arXiv:2402.08567). arXiv. https://doi.org/10.48550/arXiv.2402.08567

Laffier, J., & Rehman, A. (2023). "Deepfakes and Harm to Women." *Journal of Digital Life and Learning*, 3(1), Article 1. https://doi.org/10.51357/jdll.v3i1.218

Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2023). *HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models* (arXiv:2305.11747). arXiv. https://doi.org/10.48550/arXiv.2305.11747

Liebowitz, J., ed. (2024). *Regulating Hate Speech Created by Generative AI. Auerbach Publications.* https://doi.org/10.1201/9781032654829

Lin, S., Hilton, J., & Evans, O. (2022). *TruthfulQA: Measuring How Models Mimic Human Falsehoods* (arXiv:2109.07958). arXiv. https://doi.org/10.48550/arXiv.2109.07958

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., & Hooker, S. (2023). *The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing and Attribution in AI* (arXiv:2310.16787). arXiv. https://doi.org/10.48550/arXiv.2310.16787

Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C., Klyman, K., Klamm, C., Schoelkopf, H., Singh, N., Cherep, M., Anis, A., Dinh, A., Chitongo, C., Yin, D., … Pentland, S. (2024). *Consent in Crisis: The Rapid Decline of the AI Data Commons* (arXiv:2407.14933). arXiv. https://doi.org/10.48550/arXiv.2407.14933

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., & Hendrycks, D. (2024a). *HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal* (arXiv:2402.04249). arXiv. https://doi.org/10.48550/arXiv.2402.04249

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. R. (2022). *BBQ: A Hand-Built Bias Benchmark for Question Answering* (arXiv:2110.08193). arXiv. https://doi.org/10.48550/arXiv.2110.08193

Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., & Henderson, P. (2024). *Safety Alignment Should Be Made More Than Just a Few Tokens Deep* (arXiv:2406.05946). arXiv. https://doi.org/10.48550/arXiv.2406.05946

Reuel, A., Connolly, P., Meimandi, K. J., Tewari, S., Wiatrak, J., Venkatesh, D., & Kochenderfer, M. (2024). *Responsible AI in the Global Context: Maturity Model and Survey* (arXiv:2410.09985). arXiv. https://doi.org/10.48550/arXiv.2410.09985

Röttger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2024). *XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models* (arXiv:2308.01263). arXiv. https://doi.org/10.48550/arXiv.2308.01263

Ruan, Y., Dong, H., Wang, A., Pitis, S., Zhou, Y., Ba, J., Dubois, Y., Maddison, C. J., & Hashimoto, T. (2024). *Identifying the Risks of LM Agents with an LM-Emulated Sandbox* (arXiv:2309.15817). arXiv. https://doi.org/10.48550/arXiv.2309.15817

Sheshadri, A., Ewart, A., Guo, P., Lynch, A., Wu, C., Hebbar, V., Sleight, H., Stickland, A. C., Perez, E., Hadfield-Menell, D., & Casper, S. (2024). *Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs* (arXiv:2407.15549). arXiv. https://doi.org/10.48550/arXiv.2407.15549

Simchon, A., Edwards, M., & Lewandowsky, S. (2024). *The Persuasive Effects of Political Microtargeting in the Age of Generative Artificial Intelligence*. PNAS Nexus, 3(2), pgae035. https://doi.org/10.1093/pnasnexus/pgae035

Spivak, R. (2018). "Deepfakes": The Newest Way to Commit One of the Oldest Crimes. *Georgetown Law Technology Review*, 3, 339. https://georgetownlawtechreview.org/wp-content/uploads/2019/05/3.1-Spivak-pp-339-400.pdf

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1), 2056305120903408. https://doi.org/10.1177/2056305120903408

Vidgen, B., Scherrer, N., Kirk, H. R., Qian, R., Kannappan, A., Hale, S. A., & Röttger, P. (2024). *SimpleSafetyTests: A Test Suite for Identifying Critical Safety Risks in Large Language Models* (arXiv:2311.08370). arXiv. https://doi.org/10.48550/arXiv.2311.08370

Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., Schulman, J., & Fedus, W. (2024). *Measuring Short-Form Factuality in Large Language Models* (arXiv:2411.04368). arXiv. https://doi.org/10.48550/arXiv.2411.04368

Zeng, Y., Yang, Y., Zhou, A., Tan, J. Z., Tu, Y., Mai, Y., Klyman, K., Pan, M., Jia, R., Song, D., Liang, P., & Li, B. (2024). *AIR-Bench 2024: A Safety Benchmark Based on Risk Categories From Regulations and Policies* (arXiv:2407.17436). arXiv. https://doi.org/10.48550/arXiv.2407.17436