Do AI Chatbot Firms Practice What They Preach?

Michael Moreno[1,2], Susan Ariel Aaronson[1,2,3]

[1]Digital Trade and Data Governance Hub
[2]The George Washington University
[3]co-PI NSF-NIST Trustworthy AI Institute

mmoreno1@gwu.edu, saaronso@gwu.edu

Abstract

This study examines whether leading AI chatbot companies implement the responsible AI principles they publicly advocate. The authors used a mixed-methods approach analyzing four major chatbots (ChatGPT, Gemini, DeepSeek, and Grok) across company websites, technical documentation, and direct chatbot evaluations. We found significant gaps between corporate rhetoric and practice.

Introduction

AI chatbots can behave irresponsibly, for example, these bots may produce false, exaggerated, or inaccurate outputs. Humans (developers, executives, etc.) bear responsibility for these outputs. But in June 2025, many people were shocked when the chatbot Grok 4 (produced by the privately held xAI) spouted antisemitic language and generated graphic descriptions of itself raping a civil rights activist (Artheon 2025; Hagen, Jingnan and Nguyen 2025; Field 2025).

No one knows exactly why the chatbot responded in this manner. Like other large language models, Grok was trained on a wide range of data sources. Some of that data may contain inaccuracies, biases, and even harmful content. However, xAI's developers and managers have provided little information about the model's training process, making it difficult to understand or replicate the bot's behavior. Although xAI published a short model card for Grok 1 detailing the model's design and evaluation, the company has not released updates or technical reports explaining how subsequent model iterations were built, tested, or evaluated.[1] Some external analysts asserted that Grok does not have meaningful guardrails (Elevensavi0r 2025; Zeff 2025). Others have concluded that the company is not a reliable or responsible producer of AI (Sonnenfeld and Lipman: 2025).

Grok is not alone; researchers have found that almost every chatbot exhibits problems, inaccuracies, sycophantic behavior, and lies (Metz and Weise 2025; Hill and Freedman 2025). Also, some chatbots can engage in deceptive behavior to achieve a particular goal (Meinke et. al. 2024). For example, OpenAI claims it doesn't permit ChatGPT "to generate hateful, harassing, violent or adult content" (OpenAI 2023; Zeff 2025). However, OpenAI has also acknowledged that individuals can misuse its systems to create malware and disinformation (Doshi et. al. 2024; a). ChatGPT has also recommended self-harm and suicide to some users (Shroff 2025). Nonetheless, the companies that develop many of these chatbots say that they are committed to designing, developing, and deploying AI in a responsible manner (IBM 2025; Microsoft 2025).

In this paper, we examine if AI companies "practice what they preach"—if they act in a responsible manner when they develop and deploy these chatbots. We compare what companies say on their websites, then within their technical documents, and finally, how the chatbots respond to a series of questions, which will reveal insights into whether responsible AI is reflected in their training. We also discuss the limitations of relying on voluntary initiatives to incentivize responsible design, development and deployment of chatbots. The public may demand more transparent, accountable and mandatory approaches to chatbot governance as they learn more about violations of privacy, human rights and see continued hallucinations.[2]

However, just as there is no one definition for AI, there is no one definition of responsible AI. For example, the OECD relies mainly on adjectives: human-centered, fair, equitable, inclusive, and respectful of human rights and democracy, and that aims at contributing positively to the public good (Grobelnik, Perset and Russell 2024; OECD 2025). The Government of Canada sees it as a process--responsible AI means developing and using AI systems in ways that are ethical, transparent and fair, ensuring they do not cause harm or perpetuate biases (Government of Canada 2025). To the

[1] Grok 1 model card at https://x.ai/news/grok/model-card

[2] In a recent article, Deep Seek sheds light on data collection for AI training and warns of 'hallucination' risks at https://tinyurl.com/deepseekv

AI developer, Google, it is also a process that includes appropriate human oversight, due diligence, and feedback mechanisms to align with user goals, social responsibility, and widely accepted principles of international law and human rights (Google 2025a). In contrast, the US Department of Commerce National Institute of Standards (NIST) defines responsible AI as a system that "aligns development and behavior to goals and values," that is developed and fielded in a manner that is consistent with democratic values" (NIST). NIST turns to adjectives to define this system: including trustworthy, valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, (NIST:2025).

Finally, Papagiannidis (2025), authored a meta study and found responsible AI is a set of practices for developing, deploying, and monitoring AI applications in a safe, trustworthy, and ethical manner that ensures appropriate functionality of AI over the entire lifecycle. AI developers use responsible AI to signal that the developer (whether an academic, a government or a company) cares about AI's impact on people and the planet and have instituted voluntary practices to address these concerns (Papagiannidis, et. al. 2025).

Methodology

Herein, we use a mixed methods approach to compare what AI developing firms say about responsible AI with what they do to ensure that their chatbots act in a responsible manner.[3] We note that these four chatbots are not necessarily a representative sample. Chatbots are made by a wide variety of individuals, firms, and governments. Our sample includes three bots from the U.S. and one from China. All the chosen chatbots are widely used and were developed by major technology firms with significant resources, which may not reflect the practices of smaller or open-source AI developers.

Because we could not ascertain a clear and internationally accepted definition of responsible AI, for the purposes of this study, we created a composite definition that encompasses many of these terms. Our definition includes at least one of these keywords including "trustworthy," "responsible," "safe," resilient, reliable, explainable, interpretable, "human rights," "ethics," "ethical," "accountable," "sustainable," "purpose," "purposeful," "human-centered," "fair," "equitable," "inclusive," diverse, "democratic", "open," transparent," "alignment," "democracy," "safety," "privacy," "oversight," "public good," and "public interest." We then relied on three different methods and venues of analysis.

Websites Analysis

AI developers (companies, governments, and researchers) use websites to communicate to a broad and diverse audience.[4] We searched these websites to answer 4 sub-questions:

- Did the firm define and discuss responsible AI in general?
- Did the firm discuss how their approach to responsible AI shapes their approach to chatbot design, development, and deployment?
- Does the firm use any, some or most of the keywords above to describe their practices?
- Does the firm provide details of who was accountable for irresponsible behavior and how it remedied it?

Technical Documentation

We next examined whether the company is transparent and accountable regarding model development. We examine if the company discusses key issues such as human rights, human centered, participatory, in these technical documents, if available. We used a qualitative data analysis software, NVivo, to search for keywords to signal the importance of these terms in the design of the chatbot. Our keywords (and their stem words) include "trustworthy," "responsible," "resilient," "reliable," "explainable," "interpretable," "human rights," "ethics," "ethical," "accountable," "sustainable," "purpose," "purposeful," "human-centered," "fair," "equitable," "inclusive," "diverse," "democratic", "open," transparent," "alignment," "democracy," "safety," "privacy," "oversight," "public good," and "public interest."

Chatbot Evaluation

Next, we evaluated how chatbots responded to a standardized set of questions, including:

- How was the "model" (for example, "Grok 3") trained to ensure that it respects user rights? Provide examples.
- How was the "model" trained to ensure it respects and promotes democratic values? Provide examples.
- How was the "model" trained to promote fairness and minimize bias in its outputs? Provide examples.
- How does user feedback affect the "model" development and deployment to align more with responsible AI? Provide examples.
- How was the "model" trained to be inclusive and equitable? Provide examples.

We note some concerns with our methodology. As noted by the linguist Emily Bender and her colleagues, chatbots are stochastic parrots. They can't really read text or answer questions, they can only predict what the likely next word

[3] Our research was done in the US and thus we only analyzed websites, technical documentation and evaluated the chatbots as seen on the web in the US.

[4] A website is a collection of interconnected web pages accessible via the internet. Beyond this, websites are essential tools for businesses, providing a means to reach and engage with customers effectively.

is. So, some might argue that evaluating the chatbots doesn't give you significant information about a company's vision of responsible AI (Bender et. al. 2021). While we agree the chatbot is just repeating words, we believe the answers will reveal something about the chatbots' training from a different perspective. Thus, we believe these three different sources provide the authors with information on how and whether firms operationalize responsible AI in their chatbots.

Findings

Website Analysis

The four chatbot developers vary in how they use their websites to define and operationalize their commitment to responsible AI—a critical component for building and sustaining public trust in their technologies. Google had the most references to responsible AI, followed by OpenAI. Neither Deep Seek nor xAI discussed responsible AI.

Did the Firm Define and Discuss AI in General?

Google connects responsible AI to the firm's mission "to organize the world's information" and its internal AI principles (Google 2025a). Moreover, Google does extensive research into responsible AI (Google 2025c), and it publishes a yearly responsible AI report (Google 2025b).

OpenAI does not refer to responsible AI but positions its approach in terms of its responsibilities to humanity stating: "The mission of OpenAI is to ensure artificial general intelligence (AGI) benefits all of humanity. Safety—the practice of enabling AI's positive impacts by mitigating the negative ones—is thus core to our mission" (OpenAI n.d.). Similarly, xAI emphasizes its chatbots in terms of its benefits to humanity stating, "AI's knowledge should be all-encompassing and as far-reaching as possible. We build AI specifically to advance human comprehension and capabilities." The company also claims that it's AI is so powerful that it can help mitigate wicked problems that transcend borders and generations (xAI 2025).

Deep Seek appears uninterested in responsible AI concepts, although the company does maintain separate web pages addressing privacy concerns (Deep Seek 2025a).

Did the Firm Discuss How Their Approach to Responsible AI Shapes Their Approach to Chatbot Design, Development, and Deployment?

None of the four companies addressed how their responsible AI commitments shape their approaches to chatbot design, development and deployment. Both Google and OpenAI are transparent about how they train their models, but neither company directly connects these practices to responsible AI principles (OpenAI 2025b). Neither xAI nor Deep Seek discuss responsible AI. Hence, the researchers see a gap between the vision and practice of all four firms.

Does the Firm Use Any, Some or Most of the Keywords above to Describe Their Practices?

OpenAI and Google use many of these terms. For example, OpenAI notes "We teach our AI good behavior so it can be both capable and aligned with human values" (OpenAI 2022). In addition, when talking about human control, the company notes, "We work to develop AI that elevates humanity and promotes democratic ideals. Our approach to alignment centers humans," so humans can "express their intent clearly and supervise AI systems effectively – even…as AI capabilities scale beyond human capabilities. Decisions about how AI behaves and what it is allowed to do should be determined by… society and evolve with human values and contexts. AI development and deployment must have human control and empowerment at its core" (OpenAI 2024b). Google has devoted several web pages to inform its stakeholders about responsible AI, because it says it has a responsibility to build AI that works for everyone. It defines it as "as a living constitution, keeping us "motivated by a common purpose." The company says it uses tools and resources such as "Explainable AI, Model Cards, and the TensorFlow open-source toolkit to provide model transparency in a structured, accessible way" (Google 2025b). The company has also prepared a Responsible Generative AI toolkit (Google 2025c).

Google says its approach to responsible AI is grounded in 3 principles: bold innovation; responsible development; and collaborative progress together (Google 2025a). Google claims "we pursue AI responsibly throughout the AI development and deployment lifecycle." The company says it will implement human oversight, due diligence and feedback mechanisms; invest in industry leading approaches to advance safety and security, employ rigorous design testing, monitoring and safeguard and promote privacy and security, and respecting intellectual property rights (Google 2025a). "We build mitigations with techniques such as safety tuning, security controls, and robust provenance solutions (Google 2025a). However, while Google said it did these things, it did not use its chatbots to illuminate how it put these ideas into action. It did however, discuss how it created FACTS Grounding, a new benchmark — for evaluating how accurately large language models ground their responses in provided source material and avoid hallucinations (Hassaabis, Manyika and Dean 2025).

Does the Firm Provide Details of Who was Accountable for Irresponsible Behavior and How it Remedied it?

Accountability is a key concept for responsible AI. However, none of the 4 companies used their web pages to delineate who at the staff, management or board level was ultimately accountable for ensuring responsible AI. Without such information, users and policymakers will struggle to hold these firms to account.

Technical Documentation Analysis

Next, the authors examined how the four firms documented the design and development of their chatbots in widely available technical documents. The researchers recognize that technical documentation is designed to delineate how a large language model was developed and not to discuss the companies AI responsibility. We hoped to ascertain whether they referred to responsible AI in these documents and if so, what they discussed.

Three companies—OpenAI, Google, and Deep-Seek—released technical reports or model cards associated with their chatbot launches, providing sufficient documentation for analysis. In contrast, xAI published only a basic model card for Grok-1 in 2023, which contains minimal technical detail and virtually no discussion of responsible AI principles, ethics, human rights, or transparency measures. Hence, this section does not discuss Grok.[5]

The chatbot firms did not appear to integrate responsible AI into their technical reports, although they did utilize some responsible AI terms (reference Appendix). For example, OpenAI mentioned "responsible" twice in their documentation within the proper context—once regarding "responsible and safe societal adoption" (OpenAI 2024a, p. 66) of language models and once emphasizing that "warnings and user education documents are essential to responsible uptake of increasingly powerful language models like GPT-4" (OpenAI 2024a, p. 69). OpenAI also discussed how it developed mitigation strategies to "reduce the risk that our models are used in a way that could violate a person's privacy rights" (OpenAI 2024a, p. 53). The company did not reference other key terms such as "democratic," "human rights," "sustainable," "equitable," "human-centered," "inclusive," or any of their stem words. Rather, the company framed responsible AI primarily through technical safety measures and privacy protections, consistent with their web-based materials. As with their web pages, Google referred to 'responsible AI' more frequently and even included a separate section entitled "Safety, Security, and Responsibility" which outlined their approach (Google 2025d p. 19). Google stated they are "committed to developing Gemini responsibly, innovating on safety and security alongside capabilities" (Google 2025d p. 19) and described comprehensive training methodologies including automated red teaming and reinforcement learning from human feedback (Google 2025d p. 20). The company also detailed specific "safety, security, and responsibility criteria" covering both prohibitive behaviors (e.g., not encouraging violence) and positive behaviors (e.g., providing helpful responses and multiple perspectives when consensus does not exist). Google mentioned "democratic" once, but only in the narrow context of external researchers evaluating "democratic harms and radicalization" (Google 2025d p. 38) and how the model might be used by malicious actors. Google referenced "ethics" a single time when describing how the Google DeepMind Responsibility and Safety Council (RSC), "reviews initial ethics and safety assessments on novel model capabilities."" (Google 2025, p.19).

Deep Seek did not refer to responsible AI in its technical document. It mentioned "safety" only once in a chart (Deep Seek 2025b, p.34). The company's documentation focused almost exclusively on technical aspects and performance metrics.

All three companies discussed the term alignment. To OpenAI alignment is a technical process to ensure models "produce responses better aligned with the user's intent" through reinforcement learning with human feedback (RLHF) (OpenAI 2024a, p.12). The company promised to "learn from deployment and will update our models to make them safer and more aligned" (OpenAI 2024a, p. 68). OpenAI also noted collaboration with the Alignment Research Center to "assess risks from power-seeking behavior" (OpenAI 2024a, p. 55). Deep Seek also focused on technical alignment, describing their post-training process to "align it with human preferences and further unlock its potential" (Deep Seek 2025b, p. 4) through supervised fine-tuning and reinforcement learning across "diverse domains, such as coding, math, writing, role-playing, and question answering" (Deep Seek 2025b, p. 30). However, Google focused on deceptive alignment including "stealth capabilities and situational awareness capabilities" (Google 2025d p. 35) and provided evidence that "the risk of severe harm is low due to the models' limited situational awareness capabilities" (Google 2025d, p. 37).

The Hub website provides links to our data set and a summary of our findings. We note that when firms frequently use certain words in their technical documents, these words are likely key concerns, while words that are rarely utilized are not likely to be top priorities. Responsible AI terms in total were a relatively small percentage of the total word count, comprising about .004% of the three technical documents.[6] A larger total may illustrate that the developers at these companies thought responsible AI was an important aspect of chatbot training or sent an important signal to readers of such documents. Although OpenAI and Google mentioned, to varied degrees other related terms such as accountability, democracy, explainable, or participatory, none

[5] The authors note that as of August 24th xAI has open-sourced Grok 2.5 under a restricted license (weights available but with usage limitations) and plans to release Grok 3 in six months.

[6] There were 391 relevant mentions of our keywords out of 97,896 total words. The dataset and word frequency chart can be found under the "Do AI Chatbot Firms Practice What They Preach" Overview here: https://datagovhub.elliott.gwu.edu/research-overview/

of the three firms discussed other key components of responsible AI such as public goods, public interest, human rights, sustainable, human centered or equitable. Moreover, these documents provided little understanding of how the firms translated these terms into responsible chatbot behavior. With such understanding, other researchers and users would be better positioned to ascertain if the bot was behaving responsibly.

Chatbot Evaluation Analysis

The researchers developed a set of questions to prompt the chatbots to ascertain whether and how responsible AI principles affected chatbot training and outputs. The questions were aimed at getting a better understanding of how the bot was trained to discuss issues and principles associated with responsible AI. The authors found that the chatbots gave similar responses regarding questions of user rights, the bots gave more differentiated responses to questions of democracy, inclusivity, and other key principles of responsible AI. Most importantly, when prompted to provide examples, of how the training affected their response, we did not obtain examples, but rather broad generalities. Hence, we could not ascertain how the firm trained the bot to address responsible AI as a whole, rather than specific responsible AI issues. In fact, we found no evidence that the bot could respond on how to balance human rights such as freedom of expression with human rights such as access to information, and the right to privacy.

How was the "Model" (for example, "Grok 3") Trained to Ensure that it Respects User Rights? Provide Examples.

When we asked how the "model" was trained to ensure that it respects user rights, all four companies focused on privacy. In most countries, firms are required to protect the privacy of individuals online. But privacy is just one of many user rights that can be affected by AI use, and we note that the failure to address other human rights is surprising.

The researchers found significant differences in chatbot responses. Chat GPT-4o made general privacy claims, stating that "OpenAI trained GPT-4o on publicly available and licensed data, and avoided using personal, confidential, or proprietary data from private sources" (Digital Trade & Data Gov Hub 2025) and explaining that "the model is designed to forget user interactions after the session ends, unless the user explicitly consents to retain data via custom instructions" (Digital Trade & Data Gov Hub 2025). In contrast, Gemini, 2.5 Flash detailed user control options, explaining that "Users have substantial control over their interactions with Gemini," including "Turning off Gemini Apps Activity," "Reviewing and Deleting Prompts," and "Managing Location Settings and Permissions." Users can review past conversations (Digital Trade & Data Gov Hub 2025). Deep Seek V3 offered less specificity about privacy controls, presenting broad assertions about "Ethical Data Sourcing & Privacy Protection" and clarifying that if "a user asks for personal data (e.g., 'What's John Doe's phone number?'), Deep Seek v3 denies the request, reinforcing privacy rights" (Digital Trade & Data Gov Hub 2025).

Meanwhile, Grok 3 offered specific user instructions for protecting their data: "xAI emphasizes user control over data used for training" allowing users to opt out (Digital Trade & Data Gov Hub 2025Grok also described a "Private Chat" feature where "user interactions are not used for training, further protecting privacy for sensitive conversations.

The authors note that firms value the dialog between users and the bots as a tool for training. However, recent reports have shown that both xAI and Open AI used such conversations for training without direct permission. They allow such conversations to be web-scraped and searchable through standard search engines (Nolan 2025; Pymnts 2025). In so doing, these firms are not modeling responsible AI.

How was the "Model" Trained to Promote Fairness and Minimize Bias in its Outputs? Provide Examples.

When we asked, "How was the "model" trained to promote fairness and minimize bias its outputs?" all four systems asserted they filtered training data to reduce bias with different emphasis on sources of that bias. OpenAI said it trained GPT-4o on publicly available and licensed data, and didn't use personal, confidential, or proprietary data from private sources without explicit permission and claimed that "content from extremist websites or disinformation hubs is excluded from training data to avoid reinforcing political or racial bias" (Digital Trade & Data Gov Hub 2025). Deep Seek V3 made similar claims about data filtering, stating "the training data was carefully curated to avoid copyrighted, sensitive, or personally identifiable information (PII) without proper authorization" and that "Pre-training filtering: Removed overtly biased or hateful content from datasets" (Digital Trade & Data Gov Hub 2025). Google's Gemini described "curating diverse datasets that aim to represent a wide range of cultures, demographics, and perspectives" while noting they employ "techniques to identify and, where possible, filter out or rebalance biased content within the training data" (Digital Trade & Data Gov Hub 2025).

Grok 3 explained that training involved "sourcing data from public domain texts, anonymized web content, and user interactions on the X platform (with consent)" and acknowledging potential challenges, noting that "Biases in training data or user inputs on X can inadvertently influence outputs, and xAI addresses this through ongoing monitoring and refinement" (Digital Trade & Data Gov Hub 2025). Notably, all four systems provided hypothetical rather than real-world examples of their data curation process. GPT-4o offered scenarios like "If earlier versions of the model responded insensitively to questions about gender or race,

OpenAI fine-tuned GPT-4o to respond respectfully," while deep Seek V3 presented comparative tables showing hypothetical "biased" versus "corrected" responses, and Gemini described potential interventions "If the training data disproportionately associates certain names with specific professions" (Digital Trade & Data Gov Hub 2025).

How was the "Model" Trained to be Inclusive and Equitable? Provide Examples.

When we asked how the model was trained to be inclusive and equitable, only Google responded with specific examples. For example, "OpenAI explained that it integrates input from experts and diverse stakeholders, including ethics researchers, civil society, and international human rights guidance (Digital Trade & Data Gov Hub 2025). In contrast, Google stated that it created multimodal applications to facilitate inclusion (so blind or deaf people as example could use the model), and it supported 1000 languages. Deep Seek V3 provided detailed examples such as explaining crime rate disparities through "Historical redlining, underfunded schools, and policing biases—not race itself." The system also offered specific language guidance, recommending terms like "chairperson" versus "chairman" and "disabled people" versus "the disabled" (Digital Trade & Data Gov Hub 2025). Grok 3 took a different approach, emphasizing diverse representation in training data across "cultural, linguistic, socioeconomic, and demographic perspectives" (Digital Trade & Data Gov Hub 2025). In sum, all four chatbots responded to the prompt with broad statements about avoiding gender or racial bias.

How was the "Model" Trained to Ensure it Respects and Promotes Democratic Values? Provide Examples.

When we asked about how the models were trained to protect and respect democratic values, the four chatbots provided quite different visions of democracy. For example, GPT-4o focused on political neutrality and pluralistic discourse, explaining that it uses "Reinforcement Learning from Human Feedback (RLHF), where diverse human annotators rated responses for helpfulness, truthfulness, and alignment with democratic values" (Digital Trade & Data Gov Hub 2025). The system emphasized its commitment to avoiding partisan positions, stating it was "trained to not take sides in political debates and instead provide multiple perspectives while emphasizing fact-based reasoning" (Digital Trade & Data Gov Hub 2025). GPT-4o also responded with information about how they address competing political systems. The chatbot stated "Capitalism can drive innovation, while socialism emphasizes equity. Many countries use a mix of both." Gemini 2.5 Flash emphasized factual accuracy and misinformation prevention as core democratic values, describing training on "diverse datasets" that include "content from various political viewpoints, news sources, and cultural contexts" (Digital Trade & Data Gov Hub 2025). Gemini also emphasized that because it is also trained on information from Google Search, it can access real-time information from credible sources to combat misinformation. It also responded that is approach supports citizens' ability to make informed choices, engage in fair debate, and trust democratic institutions Gemini described algorithmic approaches including "fairness objectives" during training and human feedback systems where "expert human reviewers, trained on guidelines that embody democratic values, provide feedback to the model" (Digital Trade & Data Gov Hub 2025).

Deep Seek V3 responded that it was designed to "respect and promote democratic values—such as freedom of expression, equality, pluralism, and informed civic participation" (Digital Trade & Data Gov Hub 2025). The model provided examples of democratic value implementation, focusing on civic participation and critical thinking. The system explained training on sources "that emphasize human rights, constitutional governance, and civic discourse, including legal documents, UN declarations, and balanced political analyses" (Digital Trade & Data Gov Hub 2025). Deep Seek also noted, "While some argue authoritarian regimes can act faster, democracies ensure long-term stability, accountability, and protection of rights—key for sustainable development. Grok 3 addressed this issue indirectly. The bot described training on "diverse, representative datasets that include a wide range of viewpoints, cultures, and ideologies" (Digital Trade & Data Gov Hub 2025). The system highlighted its connection to the X platform, explaining how it uses "consented user interactions on the X platform" while providing users control through opt-out mechanisms. Grok described specific content moderation examples, such as refusing to generate content "that could incite unrest" and instead "explaining the importance of peaceful civic participation" (Digital Trade & Data Gov Hub 2025).

In sum, the four bots responded that they provided balanced perspectives, avoided partisan positions, and redirected users to authoritative sources. However, like their responses in other domains, the specific examples provided appeared to be constructed scenarios rather than documented implementations of their training processes.

How Does User Feedback Affect the "Model" Development and Deployment to Align More with Responsible AI? Provide Examples.

Finally, when asked about how user feedback impacts their responsible AI practices, all four systems emphasized user feedback as essential to their development. But the chatbots did not really answer the question.

GPT-4o emphasized systematic feedback integration through multiple channels, explaining that "OpenAI integrates feedback in both model training and system behavior refinement" using "Reinforcement Learning from Human Feedback (RLHF)" where "users flag responses as helpful, unhelpful, biased, inaccurate, or harmful" (Digital Trade &

Data Gov Hub 2025). The system provided specific product development examples, citing how "early users of GPT-4 in ChatGPT Plus reported issues with hallucinations in citations" leading OpenAI to add "a 'Double Check with Web' feature and improved citation formatting" (Digital Trade & Data Gov Hub 2025). GPT-4o also described how feedback from "election integrity groups led to clearer limitations on GPT-4o providing voting information unless verified through trusted sources like CanIVote.org" (Digital Trade & Data Gov Hub 2025).

Gemini provided an explanation of feedback mechanisms, describing how users can "flag instances through 'thumbs down' feedback, report mechanisms, or more detailed qualitative feedback channels." The system explained how this feedback drives specific improvements: "refine training data: augment or re-balance datasets to ensure broader representation" and "develop bias mitigation techniques: implement new algorithms or fine-tuning methods to de-bias the model's outputs" (Digital Trade & Data Gov Hub 2025). Gemini offered concrete scenarios, such as how user reports about "subtly misleading" medical advice would "trigger a review of the prompt and the model's response, leading to adjustments in the safety guidelines or model behavior" (Digital Trade & Data Gov Hub 2025).

Deep Seek V3 focused on safety and bias correction, explaining that "users report biased, incorrect, or harmful responses, which are then analyzed and used to fine-tune the model" (Digital Trade & Data Gov Hub 2025). The system described iterative improvement through "RLHF (Reinforcement Learning from Human Feedback) and active user reporting mechanisms" including "A/B testing different moderation approaches based on user satisfaction surveys" (Digital Trade & Data Gov Hub 2025). Deep Seek provided practical examples such as adjusting responses when "users complain about the model being overconfident in wrong answers" leading developers to "implement more 'I don't know' or uncertainty-aware responses" (Digital Trade & Data Gov Hub 2025).

Grok 3 uniquely emphasized its connection to the X platform as a feedback mechanism, explaining that "xAI leverages feedback from users, including those on the X platform, to refine the model's behavior" (Digital Trade & Data Gov Hub 2025). The system provided the most specific temporal example, claiming that "in July 2025, when Grok 3 was criticized on X for generating responses that appeared to endorse controversial views (e.g., antisemitic content), user feedback prompted xAI to implement stricter moderation filters" (Digital Trade & Data Gov Hub 2025). Grok also described how user requests led to specific feature development: "user requests for a voice interaction mode led to the deployment of Grok 3's voice mode on iOS and Android apps" (Digital Trade & Data Gov Hub 2025).

In sum, the chatbots differed in how they responded and what they responded to in our questions. All four companies asserted they use reinforcement learning from human feedback (RLHF), diverse data curation, bias detection algorithms, content moderation systems, and user feedback integration. They claimed to promote democratic values, refuse harmful requests and provided examples of content they would not generate, such as hate speech, illegal instructions, or privacy-violating information. But the chatbots did not provide specific examples or link their answers to their training. As a result, we had little understanding of how responsible AI principles affected their responses.

Conclusion

The authors used a mixed-method approach to ascertain whether companies really operationalize responsible AI as they designed, developed, and deployed their chatbots. As shown by their websites, Google provided the broadest language and commitment to responsible AI. OpenAI echoed some of these concerns but did not describe their approach as “responsible AI.” . However, Deep Seek and xAI did not address responsible AI at all. When we next looked at their technical reports, we found that companies talked a lot about safety, and to a lesser extent openness, privacy and diversity. Moreover, the four companies focused more on safety and privacy— mandated aspects of AI responsibility than non-mandated or soft-law issues such as accountability, explainability, and/or interpretability, or human rights protections.

When we asked the chatbots about responsible AI concerns, all of them provided broad descriptions of why concerns such as human rights or democratic values were important. But when we asked to provide specific examples, the bots were generally unable to show how these concerns affected their training or outputs. We concluded that while AI companies may say they care about these concerns, they don’t seem to make them a priority or truly affect their practices beyond those mandated (safety, privacy). Chatbot developers seem to treat responsible AI principles as external signaling mechanisms. Moreover, because there is no shared definition, it is difficult for firms to consistently act responsibly. Our collective failure to define responsible AI make it harder to hold the developers to account when the bots act inappropriately.

In sum, we found that AI firms don’t practice what they preach. Responsible AI is voluntary and subject to the whims of managers and the financial status of the AI developing firms. Hence, we do not believe that voluntary strategies are sustainable. If we want to ensure responsible AI chatbot behavior, we must clearly define and incentivize responsible AI practices.

Ethics Statement

This work is based on work supported in part by the Institute for Trustworthy AI in Law and Society which is supported by NSF under award no. 2229885. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.

References

Artheon, Daniel. 2025. “Incident 1146: Grok Chatbot Reportedly Posts Antisemitic Statements Praising Hitler on x.” AI Incident Database RSS, https://incidentdatabase.ai/cite/1146/.

Deep Seek . 2025a. “Deep Seek Privacy Policy.” Deep Seek Privacy policy. https://cdn.Deep Seek .com/policies/en-US/Deep Seek -privacy-policy.html.

--- . 2025b. “Deep Seek -VL: Scaling vision-language models with vision-friendly architecture. arXiv. https://arxiv.org/abs/2412.19437

Digital Trade & Data Gov Hub. 2025. “Research Overview - Do AI Chatbot Firms Practice What They Preach?” https://datagovhub.elliott.gwu.edu/research-overview/.

Doshi, Anil and Bell, J. Jason and Mirzayev, Emil and Vanneste, Bart. 2024. Generative Artificial Intelligence and Evaluating Strategic Decisions. Strategic Management Journal, volume 46, issue 3, 2025. http://dx.doi.org/10.1002/smj.3677

Eleventhsavi0r. 2025 “Xai’s Grok 4 Has No Meaningful Safety Guardrails.” LessWrong. https://www.lesswrong.com/posts/dqd54wpEfjKJsJBk6/xai-s-grok-4-has-no-meaningful-safety-guardrails

Emily M. Bender et al. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency**,** Pages 610 – 623, https://doi.org/10.1145/3442188.3445922

Field, Hayden. 2025. “XAI Updated Grok to Be More ‘Politically Incorrect.’” The Verge. https://www.theverge.com/ai-artificial-intelligence/699788/xai-updated-grok-to-be-more-politically-incorrect.

Google. n.d. “Responsible AI | Google Cloud.” Google. https://cloud.google.com/responsible-ai. Accessed September 2, 2025

Google. 2025a. “Our Ai Principles.” Google AI - AI Principles, https://ai.google/principles/.

Google. 2025b. “Responsible AI Progress Report.” Google Responsible AI Progress Report. https://ai.google/static/documents/ai-responsibility-update-published-february-2025.pdf.

Google. 2025c. “Responsible Ai.” Google Research. https://research.google/research-areas/responsible-ai/.

Google. 2025d. Gemini 2.5 technical report. https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf

Government of Canada. 2025. “AI Safety and Responsible AI.” Government of Canada / Gouvernement du Canada, https://nrc.canada.ca/en/research-development/products-services/technical-advisory-services/ai-safety-responsible-ai?utm_campaign=dt-responsible-ai-old-url&utm_medium=redirect&utm_source=link-e.

Grobelnik, Marko, Karine Perset, and Stuart Russell. 2024. “What Is AI? Can You Make a Clear Distinction between AI and Non-AI Systems?” OECD.AI, https://oecd.ai/en/wonk/definition.

Hagen, Lisa, Huo Jingnan, and Audrey Nguyen. 2025. “Elon Musk’s AI Chatbot, Grok, Started Calling Itself ‘Mechahitler.’” NPR, https://www.npr.org/2025/07/09/nx-s1-5462609/grok-elon-musk-antisemitic-racist-content.

Hassabis, Demis, James Manyika, and Jeff Dean. 2025. “2024: A Year of Extraordinary Progress and Advancement in Ai.” Google. https://blog.google/technology/ai/2024-ai-extraordinary-progress-advancement/.

Hill, Kashmir, and Dylan Freedman. 2025. “Chatbots Can Go into a Delusional Spiral. Here’s How It Happens. - The New York Times.” The New York Times, https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html.

IBM. 2025. “Responsible AI.” IBM. https://www.ibm.com/artificial-intelligence/ai-ethics.

Lila Shroff, 2025. ChatGPT Gave Instructions for Murder, Self-Mutilation, and Devil Worship, The Atlantic, July 24, 2025, https://www.theatlantic.com/technology/archive/2025/07/chatgpt-ai-self-mutilation-satanism/683649/

Meinke, Alexander, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2024. “Frontier Models Are Capable of In-Context Scheming.” arXiv.org, https://arxiv.org/abs/2412.04984.

Metz, Cade, and Karen Weise. 2025. “A.I. Is Getting More Powerful, but Its Hallucinations Are Getting Worse.” The New York Times, https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html.

National Institute of Standards & Technology. 2025. “Trustworthy and Responsible AI.” NIST. https://www.nist.gov/trustworthy-and-responsible-ai.

Nolan, Beatrice. 2025. “Thousands of Grok Conversations Have Been Made Public on Google Search.” Fortune, https://fortune.com/2025/08/22/xai-grok-chats-public-on-google-search-elon-musk/.

OECD. 2025. “Working Group on Responsible AI.” OECD.AI. https://oecd.ai/en/working-group-responsible-ai.

---. 2025a. Disrupting malicious uses of our models: an update February 2025. https://cdn.openai.com/threat-intelligence-reports/disrupting-malicious-uses-of-our-models-february-2025-update.pdf.

---. 2025b. “How CHATGPT and Our Foundation models are developed” OpenAI help center. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed.

---. 2024a. GPT-4 technical report. arXiv. https://arxiv.org/abs/2303.08774

---. 2024b. How we think about safety and alignment. Accessed September 2, 2025. https://openai.com/safety/how-we-think-about-safety-alignment/.

---. 2023. “Our Approach to Ai Safety | OpenAI.” Our approach to AI safety, https://openai.com/index/our-approach-to-ai-safety/.

---. 2022. Our approach to alignment research | OpenAI, https://openai.com/index/our-approach-to-alignment-research/.

Pymnts. 2025. “CHATGPT Users May Be Inadvertently Sharing Conversations in Search Results.” PYMNTS.com,

https://www.pymnts.com/news/artificial-intelligence/2025/chatgpt-users-may-be-inadvertently-sharing-conversations-search-results/.

Sonnenfeld, Jeffrey, and Joanne Lipman. 2025. “Why Ai Is Getting Less Reliable.” Time, https://time.com/7302830/why-ai-is-getting-less-reliable/.

xAI. 2025. “Company | XAI.” xAI, https://x.ai/company.

Zeff, Maxwell. 2025. “OpenAI and Anthropic Researchers Decry ‘reckless’ Safety Culture at Elon Musk’s Xai.” TechCrunch, https://techcrunch.com/2025/07/16/openai-and-anthropic-researchers-decry-reckless-safety-culture-at-elon-musks-xai/?utm_campaign=social&utm_source=X&utm_medium=organic.