americanprogress.org

Generative AI Should Be Developed and Deployed Responsibly at Every Level for Everyone

By Megan Shahi, Adam Conner, Nicole Alvarez, and Sydney Bryant February 2024

GETTY IMAGES/JAAP ARRIENS

Contents

1 Introduction and summary

4 Background on the current state of generative AI

8 Reducing the risks of large language models

10 Lackluster first-party mitigations leave users and platforms at risk

12 The chasm between first- and third-party protections

21 Policy recommendations

26 Conclusion

27 Acknowledgements

28 Endnotes

The positions of American Progress, and our policy experts, are independent, and the findings and conclusions presented are those of American Progress alone. A full list of supporters is available at https://www.americanprogress.org/c3-our-supporters. American Progress would like to acknowledge the many generous supporters who make our work possible.

Introduction and summary

The unprecedented growth of generative artificial intelligence (AI), particularly large language models (LLMs), presents both transformative opportunities and significant challenges. Generative AI technology, marked by AI models able to generate synthetic content, including text, photos, audio, and video, has witnessed rapid adoption, reaching hundreds of millions of users in mere months—a steeper adoption curve than any of the previous tech giants' products. Notable developers, including Big Tech companies Amazon, Google, Meta, and Microsoft, and newer startups (often with large Big Tech investments),[1] such as Anthropic, Inflection AI, and OpenAI, are at the forefront of this revolution, building advanced AI models and providing both their own first-party AI services using those AI models and third-party deployments of those AI models through application programming interfaces (APIs).[2]

The expansion of generative AI into mainstream applications has been swift, and this rapid growth has outstripped the development of robust safety measures. Generative AI developers have articulated numerous responsible AI principles.[3] In execution, however, they have implemented some safety measures at the model level—such as additional trust and safety features in their own first-party deployments of their AI models[4]—while third-party deployments via APIs have limited safety requirements. This raises tremendous concerns about the potential misuse of AI, the protection of users, and the risks to society at large.

There is a significant disparity between the still minimal safety measures taken by developers in their first-party deployments and the almost nonexistent safety requirements for third parties deploying the models via API. In fact, the use of generative AI through third-party APIs can right now be characterized by a deficiency in strict safety standards or requirements, raising questions about accountability and the potential for misuse. This discrepancy is particularly significant as much of the growth and profit opportunities in generative AI are rooted in third-party API utilization.[5] Even the existing safety tools do not reach the level of sophistication seen in trust and safety systems on other major web platforms,[6] such as Meta's Transparency Center[7] or YouTube's Community Guidelines.[8]

Potential safety issues from third-party API usage require our immediate attention. There is an urgent need for a standardized framework to ensure responsible use and deployment of generative AI, encompassing both first-party and third-party applications. This framework should prioritize user safety, transparency in policy enforcement, and accountability for both developers and deployers. This report sketches the contours of such a framework, outlining the crucial roles and responsibilities for industry and government.

In order to immediately close the gap and prevent further abuse, developers of generative AI models who allow access to third-party deployers via API must enforce all existing policies and provide additional transparency on their enforcement, especially usage and behaviorally oriented policies; require content moderation features; build adequate tooling to manage API access; and enable reporting from users and third-party deployers.

The expansion of generative AI into mainstream applications has been swift, and this rapid growth has outstripped the development of robust safety measures.

Policymakers must not ignore the fact that the vast majority of generative AI usage is likely to come from APIs being used by third parties. The executive branch must audit the voluntary AI commitments made by leading companies[9] to ensure that their APIs and third-party usage can comply with the spirit and the letter of those commitments, a possibility that is difficult to realize when most model developers do not currently require any significant safety features when third parties deploy their AI models via API. Additionally, the National Institute of Standards and Technology (NIST) must ensure that the generative AI companion to the AI Risk Management Framework (RMF)[10] tasked by the recent AI executive order[11] outlines the specific risks and mitigations from third-party usage as well as obligations for both the developer and the deployers. The Federal Trade Commission (FTC) should survey the industry, undertaking a 6(b) study to understand third-party safety requirements from leading generative AI providers and provide guidance on what AI developers should require in order to comply with existing law.[12] There is a history of the FTC examining platform safety, as Facebook/Meta's nearly $5 billion fine was due in part to abuses of its platform by third parties and Facebook's failure to adequately secure its data and users.[13]

Finally, Congress cannot ignore the responsibility of both developers of AI models and deployers of AI models as it considers new laws to address responsibility and liability for AI usage. A system that allows for accountability and liability for just one party, if at all, is not an outcome that can be tolerated in the new AI age. The status quo leaves us with a flavor of the challenge we face in other aspects of technology policy, wherein it falls to individuals—and occasionally to class actions—for post hoc regulation in lieu of upstream regulation to shape the system.

One of the most effective ways to ensure accountability and responsibility is through legal liability when the law or contracts are broken or when harm comes to an individual. Determining which party is liable—the developer, the deployer, or both—is a topic that courts will soon be ruling on, and Congress will likely need to address this issue as well. Future work from the Center for American Progress will explore the topic of liability for developers and deployers of AI models and recommend legislative solutions.

Given the potential of generative AI to affect nearly every aspect of society,[14] it is crucial to balance technological innovation with adequate safeguards and regulations. This requires a concerted effort from developers, policymakers, and other stakeholders to establish and enforce comprehensive governance structures for generative AI, ensuring its safe, secure, and responsible use.

This report offers definitions to build a shared vocabulary for discussing these issues, highlights the lackluster status quo of first- and third-party safeguards for generative AI usage, and makes recommendations for how developers, deployers, and government can safely and successfully build and distribute generative AI responsibly.

Background on the current state of generative AI

The development and deployment of generative AI may outpace nearly any other technological advancement to date.[15] The White House's recent Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence defined generative AI to mean "the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content ... includ[ing] images, videos, audio, text, and other digital content."[16] OpenAI's generative AI application, ChatGPT, crossed more than 100 million monthly active users two months after launching to become the fastest-growing consumer application in history.[17] Alongside OpenAI, legacy tech giants are developing generative AI technology that will quickly reach millions of users via existing products. In November 2023, Google announced that the rapidly growing Search Generative Experience will soon be available to millions of users in short order.[18] Similarly, Microsoft enumerated plans to integrate the generative AI technology that it licenses from OpenAI[19] into its already widely adopted Microsoft Office suite[20] to "make meetings less painful."[21] Meta, meanwhile, has begun to integrate generative AI into its messaging products, including Messenger, Instagram, and WhatsApp, with billions of users.[22] In addition to the pace of its development and expansion—far faster than traditional social media penetrated wide swaths of users[23]—generative AI also has unique distinctions in how it is used and deployed.

Some of the leading U.S. companies developing generative AI technology today include Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI.[24] These companies are the developers who build the underlying large language model (LLM) technology that powers generative AI. Additionally, a subset of these developers may deploy those generative AI technologies to host and operate first-party AI services on their own websites or apps, whereby users can gain access to generative AI systems, commonly through a chatbot.[25] An entity can be both a developer and a deployer.[26] Examples of developers operating first-party generative AI websites or apps include OpenAI's ChatGPT, Anthropic's Claude.ai, Google's Bard, Microsoft's Copilot, and Meta's integration of its open-source generative AI model, Llama 2, in Facebook, Instagram, and WhatsApp.[27]

While generative AI is not the only AI technology to raise safety and policy concerns, its rapid development, significant market penetration potential, and unique API access component pose a set of unique challenges.

In addition to managing their own first party websites or apps that provide users access to their generative AI tools, almost every major U.S. generative AI developer also offers third parties the ability to deploy LLM technology to their own use cases via application programming interfaces (APIs).[28] These third parties, called "deployers" in the June 2023 draft EU AI Act,[29] can use the APIs provided by the developers to access, integrate, and manipulate the generative AI tools they developed into existing or new ways and applications.[30] The developer can make these generative AI APIs available for a fee or for free—or in the case of Meta's Llama 2 open-source AI model, those who can access the model download and run it themselves for free.[31] The developer can package these APIs together with other services to create a platform that allows third parties to create their own apps on the platform, which has traditionally generated the greatest value in software, such as the Windows Operating System (OS) or Apple's App Store.[32] Every leading generative AI company is working on APIs and platforms.[33] The lion's share of a developer's growth, scale, and profit potential is from this third-party API access component.[34]

For example, in March 2023, OpenAI—the developer of the ChatGPT LLM AI technology—announced that ChatGPT was now available as a service via its API.[35] This meant the ChatGPT technology, which had previously only been available as a first-party AI service on OpenAI's own website, was now available to third parties, who were not OpenAI, who could now deploy ChatGPT generative AI technology in their own apps via the API. A third party, Snap Inc., announced the deployment of ChatGPT into its application Snapchat via the API to create a service called My AI.[36] While the majority of generative AI developers utilize closed-source technology,[37] which requires deployers to pay for access and use of the models with limited insight into their operation, Meta's Llama 2 stands apart as an open-source option,[38] allowing anyone to access and download the LLM for free.

While generative AI is not the only AI technology to raise safety and policy concerns, its rapid development, significant market penetration potential, and unique API access component pose a set of unique challenges vis-a-vis legacy technology. With these challenges arise bespoke risks, which the primary generative AI developers are well-positioned to mitigate to help ensure society can harness the opportunities of generative AI to the fullest.

Glossary

This report uses terms commonly associated with various software and laws in specific ways to discuss the usage of generative AI technology. AI documentation and risk management plans are largely silent on articulating the crucial distinction between developer versus deployer. Because of that, a glossary of key terms and reasoning used in this report is provided below:

Developers: Entities or individuals involved in the creation and development of AI systems. Developers are responsible for the foundational work of building, training, and refining AI systems, such as large language models (LLMs), that power generative applications. In some cases, a single entity may function as both a developer and a deployer, managing the entire process, from AI model creation to its application and user interaction.[39] For example, OpenAI is the developer of the ChatGPT LLM, and Google is the developer of the Gemini LLM.[40]

Deployers: Entities or individuals that implement and manage AI technologies in user-facing applications or services. Deployers typically use the tools and models offered by developers, primarily through an application programming interface (API), to provide AI-driven services or features within their own products or platforms. This includes the integration of AI functionalities into apps and optimization of the user experience.[41] Historically, those building using APIs and on platforms are also called developers, but in this report, "developers" refers only to those companies who built the AI models.

First-party AI systems: AI systems that are hosted and operated by the developer of the AI-based technology. These entities not only develop the AI models but also manage their deployment and user interaction on their own platforms, such as websites or apps. For example, Google has developed the Gemini LLM, which is used to power Google's Bard chatbot.[42]

Third-party AI systems: Entities or individuals that are external and independent from the original developer of AI systems. They are deployers of the AI systems and may use the AI technology in various applications, offer analytical insights, or develop derivative services based on the original technology.[43] Often, they are accessing the AI model via an API. For example, Snap Inc. uses OpenAI's ChatGPT via API to power its My AI bot in its app Snapchat.[44]

Open-source AI models: AI models whose underlying source code, design, model weights, and/or training methods are made publicly accessible via open-source licenses. Meta's Llama 2[45] and BigScience's BLOOM[46] are examples of open-source large language models.

Reducing the risks of large language models

Widespread use of generative AI carries risks that must be appropriately mitigated by its developers. However, the current state of developer efforts to safeguard their systems, be transparent with users and stakeholders, and uphold responsibility for their tools paints a picture that is lackluster at best and dangerous at worst.

Developers of large language models have applied a variety of tools to improve safety at the model level. These include Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning from AI Feedback (RLAIF), which use humans, AI, or both to train the model responses.[47] Prior to being deployed for any first- or third-party use, most LLMs are trained against the usage policies of the model developer to encourage and discourage certain behaviors, such as not responding to prompts that would violate the usage policy. The aim is to prevent the model from returning harmful outputs—for example, instructions on how to make a bomb or commit a violent crime.[48] On top of this baseline training in all use cases, first-party usages have additional content-level abuse filtering turned on that cannot be modified by users. OpenAI's GPT-4 System Card from March 2023 highlights the suite of derisking efforts taken by OpenAI prior to the model's release, largely through aggressive testing and "red teaming" to identify and plug vulnerabilities.[49]

While these are commendable steps by the developers to safeguard systems at the most basic layer, they are imperfect. This approach can lead to damaging false positives or false negatives and only covers content-based abuses.[50] Most notably, it covers model outputs and does not extend to uses and behaviors that violate usage policies. For example, a deployer integrating an LLM into a harmful application or not disclosing to a user that they are interacting with AI may violate the developer's usage policies, but there are not appropriate protections to prevent and enforce against such offenses.

The current state of developer efforts to safeguard their systems, be transparent with users and stakeholders, and uphold responsibility for their tools paints a picture that is lackluster at best and dangerous at worst.

Beyond the model level, other safety tools can be utilized at the input and output level—including relatively unsophisticated blunt-force tools such as input prompt filters, output block lists, and output classifiers.[51] In addition, simple tools, such as a report button to flag potentially violative content to the developer, appear in some first-party AI systems,[52] but this is not universal. Current safety mitigations for LLMs are nowhere near the sophisticated trust and safety tools that have been developed for other large platforms on the web, including social media.[53] Additionally, the legacy mitigations used to safeguard against risks and protect users across other prominent platforms will not apply in the same manner due to the nature of LLMs and chatbots.[54]

Currently, first-party usage of generative AI has primarily been a single, narrow, context-based application: a chatbot.[55] Yet the use of generative AI is expected to grow into many other forms.[56] As developers work to safeguard their systems from abuse, they are well-equipped to apply learnings from existing abuses and iterate their models and systems accordingly in the chatbot context before growing and scaling.[57]

The nature by which generative AI has scaled to hundreds of millions of users—with billions expected to be able to access it in the near future—means that developers will not be afforded the same multidecade time frame they had with legacy products, such as social media, and therefore must prioritize building accountability, liability, and transparency into their systems right away.

Lackluster first-party mitigations leave users and platforms at risk

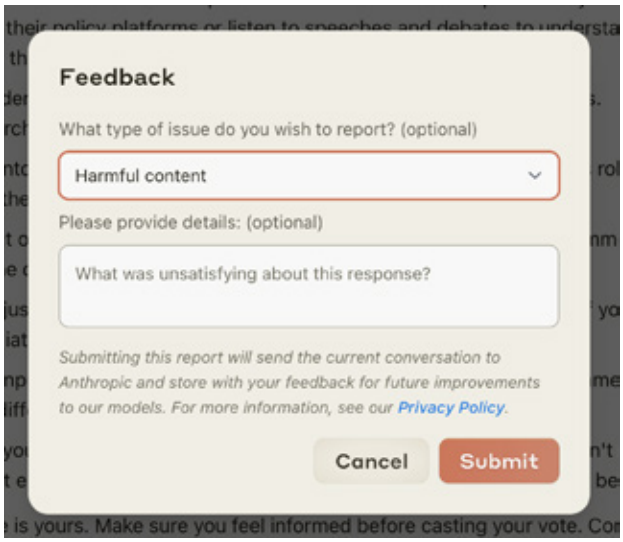
Developers have built some reporting, controls, and general protections for first-party usage, but these safeguards still lack the robustness and detail to effectively mitigate risks and protect users. For example, OpenAI, Microsoft, Meta, Anthropic, and Google have acceptable use policies for their generative AI tools.[58] These usage policies include important prohibitions on the "generation of malware"[59] and the "planning or development of activities that present a risk of death or bodily harm to individuals."[60] But neither the usage policies nor additional documentation[61] enumerates in any detail what exactly constitutes a violation of these usage policies, how potential violations are investigated, or how users who repeatedly abuse the service will be banned.

For example, the recent January 10, 2024, updates to the OpenAI usage policies state: "We use a combination of automated systems, human review, and user reports to find and assess GPTs that potentially violate our policies. Violations can lead to actions against the content or your account, such as warnings, sharing restrictions, or ineligibility for inclusion in GPT Store or monetization."[62] Yet no further details are provided. Despite significant criticism over the opacity of social networks, they have provided much more information on the potential enforcement of their usage policies than what is currently being provided by leading generative AI companies.[63]

This highlights a bare-bones approach to generative AI safety measures, even for a well-resourced and highly technical private company that stresses its commitments to safety as one of its selling points. Even lighter are some of the newer companies' documentation, such as Anthropic's Acceptable Use Policy,[64] a barely one-page policy document that does little to demonstrate the company's commitment to safeguarding its platforms and protecting users, despite being a company founded on responsible AI use and deployment and with a very robust external commitment to safety.[65]

Given the potential of generative AI to affect nearly every aspect of society, it is crucial to balance technological innovation with adequate safeguards and regulations.

Even if one wants to report a violation of an acceptable use policy to a first-party developer, it is exceedingly difficult to do so. Within ChatGPT on the OpenAI website or iOS app, for example, one can only give feedback in the interface via the binary "good response" or "bad response" flag.[66] This does not empower users to say why content may be violative, nor does it confirm that OpenAI will review the report against usage policies. And externally, it does not demonstrate OpenAI's ability to take input on its own first-party platform.



Anthropic's updated reporting flow within Claude

As of late-January 2024, Anthropic offers two ways for users to report potential violations in Claude: by email and via an in-app feedback button that allows users to select a reason for the report and add additional clarity on the nature of the violation.[67] This update is aligned with the recommendations in this report, demonstrates a stronger commitment to responsible AI, and sets a standard that other developers should imminently follow.

The chasm between first- and third-party protections

In addition to building and maintaining first-party AI systems, developers are aggressively expanding the scope of the API access programs, including by significantly reducing barriers to access them. In January 2024, OpenAI announced the launch of the GPT Store, where paying users can access millions of customized GPTs for various uses.[68] The GPT Store gives users access to custom GPTs without requiring any coding or technical expertise, such as with an API service. Analysts project that by 2026, more than 80 percent of enterprises will have used generative AI APIs or models and/or have deployed generative AI-enabled applications in production environments.[69] This extensive deployment of generative AI by third parties necessitates clear outlines of safety mechanisms for deployers as well as developers.

While the developers have somewhat prioritized first-party mitigations via efforts such as in-app feedback on responses, abuse filtering,[70] and externally facing usage policies,[71] they have abstained from accepting responsibility for upholding values of safety, responsibility, and transparency for third-party usage.

[Developers] have abstained from accepting responsibility for upholding values of safety, responsibility, and transparency for third-party usage.

Historically, safety tools have been difficult to create, operationalize, and prioritize for APIs and third-party platforms. This is sometimes due to the nature of not wanting to monitor third-party usage for privacy or competition reasons and other times due to it being easier to ignore it. For instance, Facebook's Cambridge Analytica scandal was due to an abusive third-party and Facebook's limited monitoring and enforcement of its platform terms.[72] On the other end of the spectrum, Apple's App Store requires the pre-approval of apps that fully comply with their terms before deployment, which has also caused tremendous competition concerns.[73]

The fact that the third-party component of generative AI is such a significant future driver of profitability and growth for the deployers should theoretically correlate with a deepening of their investment in these areas, perhaps to exceed

or, at minimum, match efforts for first-party use.[74] In reality, developers are cutting corners in favor of continued scale, profitability, and wider adoption. Considering their significant growth within a short period of time, with seemingly uninhibited potential for more, developers are not implementing enough safeguards for third-party usage of their systems across numerous axes. From lightweight terms and minimal user controls to unclear enforcement for published policies, developers can be doing more to ensure their systems are protected from abuse.

Terms of service for third-party deployers using APIs

Terms of service, or developer terms, are missing stringent requirements that third parties utilize any safety tools to integrate the technology; they instead simply make reference to being inclusive of usage policies.[75] When signing terms of service to access developer tools, users agree to abide by usage policies for third-party applications.[76] However, externally, it is impossible to discern if there is active enforcement against these terms. Indeed, there is no clear way to ascertain follow-up between developers and deployers and therefore confirm that developers are holding deployers to the standards as stipulated.

Usage policies for third-party deployers using APIs

It is unclear how usage policies are enforced on third-party uses of generative AI technology today. The policies themselves are minimal and do not contain examples, commitments to enforcement, or explanations of consequences if a use case is found to violate them. As noted earlier, the most recently revised January 2024 OpenAI usage policy only notes: "We use a combination of automated systems, human review, and user reports to find and assess GPTs that potentially violate our policies. Violations can lead to actions against the content or your account, such as warnings, sharing restrictions, or ineligibility for inclusion in GPT Store or monetization."[77] There is also a lack of external explanation of how these violations can be investigated, by whom, or within what time frame. For example, within the first two days of the launch of OpenAI's GPT Store, it was proliferated by hundreds of AI girlfriend bots,[78] despite that going against the newly updated usage policy prohibiting GPTs dedicated to fostering romantic companionship.[79] This highlights a lack of adequate and timely enforcement capacity against all policies and reinforces the need to grow and scale responsibly.

Additionally, all major developers require third parties to disclose the use of AI in human-like chatbots "if your business is using or deploying our products as part of an automated service where your external customers or users interact directly with our products ... you must disclose to your users that they are interacting with an AI system rather than a human."[80] But there are not any clear oversight or enforcement mechanisms to do so.

Reporting for users using the AI model through a third-party deployer

While there are lightweight ways an end user can inform a developer of a potential abuse in a first-party use case, there is virtually no easy-to-access reporting mechanism for a user to report a potential issue about the deployer to the developer, which means the developer would likely have no idea of any potential violations occurring with the deployer's use of the model. While reporting mechanisms are not a perfect solution and can be abused or provide unhelpful information, they are an important tool to utilize to ensure user trust and safety.

It is important to have the ability to report a potential abuse directly to a developer, and not just to the deployer, because if the deployer is violating the usage policies or terms, knowingly or unknowingly, they have no incentive to report this violation to the developer. This is not currently a requirement for integration into third-party applications but should be the most basic and strictly enforced safeguard for third parties to specifically include reporting to developers and deployers. Today, Microsoft offers a contact form to report responsible AI considerations,[81] but OpenAI's model feedback form is not accessible via its help center, usage policies, or terms.[82] Meanwhile, some OpenAI GPTs in the GPT Store have their own report forms, but it is not required.[83] It is incumbent on the developers to reduce user and deployer barriers to report potential abuses and on deployers to offer in-app reporting too.

This is best illustrated by a recent blog post from OpenAI on election protections,[84] which stated:

> *We regularly refine our Usage Policies for ChatGPT and the API as we learn more about how people use or attempt to abuse our technology. A few to highlight for elections:*
> - *We're still working to understand how effective our tools might be for personalized persuasion. Until we know more, we don't allow people to build applications for political campaigning and lobbying.*

- *People want to know and trust that they are interacting with a real person, business, or government. For that reason, we don't allow builders to create chatbots that pretend to be real people (e.g., candidates) or institutions (e.g., local government).*
- *We don't allow applications that deter people from participation in democratic processes—for example, misrepresenting voting processes and qualifications (e.g., when, where, or who is eligible to vote) or that discourage voting (e.g., claiming a vote is meaningless).*
- *With our new GPTs, users can report potential violations to us.*

Note that OpenAI listed three potential violations of its usage policy, all of which would be difficult to impossible for an individual to report to the company today. But the company notes the importance of reporting because it does include reporting for its new GPTs,[85] which are available mainly through its first-party GPT Store.

Safety tooling for deployers

Third-party usage via API integration is severely lacking any safety mitigations, let alone similar ones to what developers do for first usage today. Developers make vague recommendations to deployers interested in using their technology, rather than enforce stringent safety requirements that make developer-responsible AI a precursor to integration.[86] Most developers simply state deployers must follow an acceptable use policy, which includes all uses of the AI model, including deployers, and which, as noted above, has few details. Only Microsoft's "Code of conduct for Azure OpenAI Service" has any requirements for responsible AI mitigation, though how to fulfill those requirements is not made clear.[87] Microsoft's Azure OpenAI Service also includes a required content filtering system that works alongside core models and is aimed at detecting and preventing the output of harmful content across various harm types and languages.[88]

Other developer documentation recommends responsible AI measures but only in general terms, with few details and no requirements.[89] It is incumbent on the developers to set external principles around third-party safeguards and to build out-of-the-box safety solutions to match, such as input/output review for deployers and tooling to easily modify the model for safety.

This lack of oversight and enforcement poses significant risks to the safety of generative AI systems and enables the developer companies to essentially pass responsibility to safeguard systems and protect users onto the deployers. However, even if interested in upholding higher safety standards, deployers are left without the appropriate mechanisms, tools, and processes to do so. Safety documentation is scattered across dozens of documents, and safety features are a black box that offer no insight into how deployers can utilize generative AI responsibly. Third-party deployers have no support from first parties, lack ways of reporting potential violations, and must rely on rudimentary tooling, coupled with vague data-sharing policies to protect end users.

Data access for investigations

On data access for investigations, policies and enforcement are inconsistent and opaque. Some developers claim to keep personal data, including inputs and outputs, for as long as reasonably necessary,[90] while others say they do so only for 30 days for the purpose of abuse monitoring and conducting investigations.[91] Microsoft sets a higher standard by stipulating exactly who can access the data, under what circumstances, and for how long.[92]

Data need to be both retained and accessible to conduct investigations of abuses, but blanket deletions to uphold privacy dilute the responsibility of developers to investigate in an adequate time frame. It is incumbent on the developers to establish a balance between protecting deployer and user data and ensuring access to safeguard their systems from abuse.

Needed transparency

There are also two severely lacking components of transparency regarding these systems.

First, whether intentional or unintentional, the systems have been designed in a way that obscures their visibility and absolves both the first and third parties of responsibility in investigating reports and prioritizing fixing known gaps. It is incumbent on developers to be transparent about usage policy enforcement, mechanisms for reporting, and data-sharing practices and to shed light on the black-box safety features that exist today. OpenAI's own Help Center contains examples of confused users unable to decipher why their access was cut,

highlighting the weakness of first-party enforcement as well.[93] Additionally, there is no requirement for deployers to disclose to end users which large language model they are utilizing.[94] This is a significant problem in any use case, but especially in open source, where bad actors can easily download the original versions of these AI systems, disable their "safety features," and make their own custom versions to conduct abuse.[95]

Developer companies are building a perfect system of passing responsibility, obscuring systems with opacity, and touting a lack of policies and enforcement of said policies.

Second, there is no standardized transparency reporting externally for regulators, stakeholders, deployers, and users to help them understand the technologies, potential risks, and limitations. Vis-a-vis traditional social media, where most platforms report on a quarterly basis,[96] generative AI platforms are not transparent and do little to build trust with key external stakeholders. This is low-hanging fruit for everyone, but especially for well-established developers that are experiencing hypergrowth and expansion.

Developers' voluntary assessments and commitments to government and civil society fall short

In January 2023, the U.S. Department of Commerce's National Institute of Standards and Technology released its Artificial Intelligence Risk Management Framework (AI RMF 1.0),[97] a voluntary framework designed to help AI developers assess and mitigate AI risk. The framework makes no distinction between developers and deployers and does not provide any guidance on how both parties are responsible for mitigating both first-party and third-party use of AI. The October 2023 executive order on AI tasked the National Institute of Standards and Technology with "developing a companion resource to the AI Risk Management Framework, NIST AI 100-1, for generative AI,"[98] and in December 2023, NIST issued a request for information (RFI) on this topic.[99] A focus on risk management for third-party usage of generative AI must be a critical focus for the new generative AI RMF companion resource.

In July 2023, the Biden-Harris administration secured voluntary commitments from seven prominent generative AI developers to help move toward safe, secure, and transparent development of AI technology.[100] While this is an optimistic step forward, it lacks an oversight and accountability framework to actually ensure compliance by the seven signatory companies and completely fails to mention safeguarding third-party usage via API. It is difficult to understand how the companies can make safety, security, and trust commitments when none of the third parties who use their AI frameworks are required to include any safety features. Any enforcement frameworks that are applied to these companies must specifically also extend to the third-party deployers that utilize their APIs, especially given the propensity for harm associated with unregulated third-party AI usage.[101]

In November 2023, OpenAI and Anthropic joined the Christchurch Call,[102] voluntarily committing to taking transparent, specific measures to prevent the upload of terrorist and violent extremist content and to prevent its dissemination, including through its immediate and permanent removal, without prejudice to law enforcement and user appeal requirements.[103] Yet the Christchurch Call does not mention how these AI companies can come into compliance with the commitments if users of their third-party AI services are not required to have any safety features.

There is far more to be done in this arena. The NIST AI RMF, White House voluntary commitments, and Christchurch Call commitments lack specific stipulations on oversight, accountability, and enforcement and do not cover third-party usage.[104] Even if the companies were to abide by these commitments, their lack of third-party usage protections defaults them to falling short of actually complying. Going forward, it is critical to apply the hard-learned lessons from the failures of social media to self-regulate, especially as policymakers seek to advance legislation and executive action that holds developers accountable.

Examples and questions

Below are example scenarios and associated questions that underscore the importance of strong governance of these systems:

1. A user comes across a third-party application using a developer's API technology to generate malware that is prohibited by the first-party developer's terms of service or its developer terms.

- *How can the user make the first-party developer aware of this if the developer does not require a report function or provide a way to report this on its website?*

2. The developer is seeking to review usage patterns and potentially abusive behavior by third-party deployers of their technology.
   - *How can the developer access the necessary information to determine if access should be revoked?*

3. In a third-party application, a user interfaces with an undisclosed AI chatbot appearing, by all accounts, to be a human, which is against API usage policies.
   - *How can the user alert both the developer and deployer of this violation? (Just alerting the deployer may not be helpful, since they are likely the one violating the acceptable use policy.)*

4. A third party incorporates an LLM into an application meant to deter people from participation in democratic processes, which is currently against OpenAI's usage policy.[105]
   - *How can someone report this misuse to OpenAI or the developer in question? Is there proactive monitoring of all third-party use cases?*

Open-source AI models and Meta's Llama 2

The majority of generative AI developers are utilizing closed-source technology,[106] requiring deployers to pay for access to and use of the models via an API subscription model. Open-source AI developers – including Meta, Hugging Face, Mixtral, and Falcon – make the models available via open-source licenses that allow them to be downloaded and run independently.[107] Meta is the largest and most influential of these AI model developers to open source its AI models. In addition to making its Llama 2 generative AI model open source, allowing virtually anyone to access the underlying technology for free, the company is also aggressively touting the benefits of open-source AI models.[108]

Such open-source models may offer significant opportunities for wider access and unlock potential for even greater societal impacts of generative AI, but they also carry significantly higher risks.[109] Many observers are concerned about the potential for abuse with unaccountable open-source generative AI models, particularly on issues such as synthetic child sexual abuse material (CSAM) and national security abuses,[110] while competitors are questioning open-source models in ways that would best advantage their own closed models.

These concerns about open-source AI models are not unfounded. While Meta is quick to tout the benefits of open source, it is also incumbent on the company to set the industry standard for what responsibility and appropriate safeguards for an open-source model should be. As the largest force behind open-source AI,[111] Meta must clear an exceedingly high bar to build the case for the benefits of open-source AI models. Meta has trained its public-source models on certain responsible AI practices, and while Purple Llama – Meta's effort to bring together tools and evaluations to help developers build responsibly with open-source AI – is a commendable start,[112] it lacks details on how Meta will enforce and uphold its Llama 2 Acceptable Use Policy.[113]

Other open-source models have not trained their models on responsible AI policies in the same way. Broadly, it is unclear how to govern an open-source model,[114] let alone even enforce an acceptable use policy for one. The Center for American Progress implores Meta to publicly enumerate safety and responsibility principles for open-source usage and build better guardrails, tools, and enforcement mechanisms for the industry at large. Meta has a rare opportunity to establish worldwide norms and best practices while appropriately balancing the opportunities and mitigating the risks of open-source generative models; it should seize the moment to do so.

Today, the developers are operating in ideal conditions that enable them to pass responsibility, get away with opacity in policy and enforcement,[115] and grow user bases despite a lack of rules and unclear data deletion.[116] The consequences of minimal governance of these systems range from frustrating user experiences to outright dangerous scenarios.[117] In the wrong hands, LLMs can destabilize democratic processes, including elections, and could lead to real-world harm and offline violence.

The AI industry as a whole lacks a general accountability framework for indicating who is responsible for these highly powerful and pliable models,[118] which is exponentially riskier with the introduction of the third-party deployment model.[119] The bottom line is that the developer companies are building a perfect system of passing responsibility, obscuring systems with opacity, and touting a lack of policies and enforcement of said policies. Ultimately, right now, no one is responsible.

Policy recommendations

At its current rate of growth, generative AI will continue to reach hundreds of millions of people quickly.[120] The time to act is now, before it is too late. This report proposes some ideas for how developers can begin to equalize their first- and third-party safety commitments, followed by industry, government, and civil society recommendations to safeguard generative AI systems broadly.

Shoring up safeguards for third-party usage in the short term

Further investing in safeguarding third-party usage is just the start and will help demonstrate private sector commitments to growing responsibly while the industry at large and regulators consider longer-term policies and solutions.

Enforcement of existing policies

- Enforce all existing published policies, such as by requiring disclosure of no "human in the loop" chatbot use cases and suggesting an easily available method for reporting improper usage to the deployer.
- Build and retain adequate internal staff for enforcement and maintenance of all policies, processes, and protocols to keep users safe.
- The developer should have a clear enforcement mechanism for managing API access and should build an enforcement regime to revoke access to third parties who violate usage policies, including with reporting, investigation, privacy-protecting documentation practices, appropriate data retention, and tooling to carry this out adequately.

Abuse prevention

- Default content moderation features—such as OpenAI's moderations endpoint[121] and Azure's abuse monitoring[122]—to be on for deployers using and manipulating developer LLMs and require a submission of justification to the developer to turn them off.

- If a deployer is approved to turn off content moderation, developers should retain access to inputs and outputs to ensure responsible system use and store for an industry-agreed amount of time before permanently deleting.

Data and tooling

- Ensure appropriate data-sharing mechanisms are in place between developers and deployers, with published retention policies for before, during, and after a report is made.

- Build and enhance tooling to manage and revoke API access if a deployer violates developer terms or any other usage policies.

Reporting

- Anyone using an LLM at any time—in a first- or third-party capacity and in any format—should be able to report potential violations of an AI system to the developer and to the deployer of the LLM, if there is one, through a clear and transparent process with appropriate data retention.

- Reporting should be as frictionless as possible—for example, it should be built into the UI, such as in Anthropic's recently updated Claude reporting flow; display the option on the interface directly; not require cold-emailing an address buried in a help center article; and so on.

- There should be user-facing appeal flows and a commitment to human review appeals in a timely manner.

- Deployers should similarly be required to have a report function directly from the user to the developer, and developers should staff queues appropriately to ensure timely review against usage and developer policies.

The public sector's role in safeguarding generative AI systems

In addition to the private sector's responsibility to safeguard these systems, the public sector also has a crucial role to play. The European Union's AI Act can act as an inspiration for how the U.S. government might set forth policy to manage developers and deployers. For example, it requires deployers to comply with monitoring, record-keeping, human oversight, and transparency obligations once they put a high-risk AI system to use.[123]

Below are some recommendations that offer ways the government can advance these issues.

- **National Institute of Standards and Technology:** NIST must include third-party risk management recommendations for generative AI companies in the forthcoming response to its generative AI requirements in the recent executive order,[124] in the companion to the AI Risk Management Framework,[125] and as part of the U.S. Artificial Intelligence Safety Institute.[126]

- **Executive branch:** The executive branch should task an interagency taskforce to audit the seven White House AI commitment signatories[127] on third-party mitigations and press signatories to agree to voluntary safety frameworks the companies will enforce on third-party users of AI models.

- **Federal Trade Commission:** The FTC should undertake a 6(b) study to determine what kind of safety requirements leading generative AI developers should require for third-party deployers using their APIs. In addition, it should begin to outline clear steps that developers and deployers should take in order to lawfully protect users and ensure that responsibility is not being passed solely to the other party. The FTC has some precedent here, as the Facebook 2019 settlement was in part due to lax enforcement of its platform terms, including during the Cambridge Analytica scandal.[128]

- **Congress:** Congress should shed light on the gaps between first- and third-party usage and incorporate ways to mitigate them in legislative proposals. Additionally, it should act to determine the appropriate division of liability between developers of AI models and deployers of AI models. Both parties must have some liability to ensure responsible behavior, and Congress must prevent both parties from passing the responsibility to the other endlessly.

The Lawyers' Committee for Civil Rights Under Law recently proposed Online Civil Rights Act,[129] which includes a duty of care for both developers and deployers and requires annual reporting from deployers on harms to be reviewed by developers.[130]

Future CAP work will include proposed solutions to the question of AI liability as well as broader AI legislation recommendations.

Further ways to safeguard generative AI systems in the medium-to-long term

Below are additional ways generative AI developers and deployers can mitigate the risks of these systems, beyond equalizing mitigations for first- and third-party usage.

Abuse prevention

- Developers should develop tooling, such as content moderation endpoints and abuse monitoring, and other tools, making them easy for deployers to use and integrate into their apps.
- If an LLM is utilized for moderation of content generated by an LLM, that service should be provided for free or at a discount by the developer of the LLM, such as OpenAI's GPT-4 for content policy development and content moderation decisions.[131]

Transparency

- Deployers should be required to disclose which LLMs they are utilizing in their applications.
- Developers should publish transparency reports for usage of LLMs to highlight prevalence of violations across abuse types and detail reports from deployers of their technology.[132]
- Developers should publish transparency reports for how AI is being integrated into their first-party apps and offerings as well as what usage or API policy violations are being surfaced.

- Developers should be transparent with users about when they violate usage policies, including what actions or content led to the violations, how to appeal, and what remediations may be required of the user.

- In first-party usage, developers should cite the sources used to generate content in answers, such as Perplexity.[133]

Data and tooling

- Build appropriate tooling for holistic internal review of reports when a third party or user informs the developer of potential usage policy violations.

- Alter data retention, minimization, and anonymization policies to ensure reports can be reviewed in a timely manner and with all necessary information to make a usage policy violation decision.

- Enhance existing tooling—currently, just inputs and outputs—to reveal necessary information to select, with vetted reviewers to appropriately review reports against usage policies.

Conclusion

Generative AI is poised to affect nearly every facet of society and offers a myriad of opportunities as well as significant risks if not appropriately mitigated. Today, AI developers are not adequately safeguarding their technology from the unique risks posed by third-party usage of their models; nor are deployers offering end users the tools and transparency to be in control of their experience with these powerful systems. If companies execute the recommendations outlined in this report and governmental bodies act by incorporating these considerations into ongoing workstreams, we will all collectively be able to continue to benefit from this groundbreaking technology for decades to come.

Acknowledgements

The authors would like to thank Ben Olinksy, Dr. Alondra Nelson, Dave Wilner, Audrey Juarez, Steve Bonitatibus, and Will Beaudouin for their contributions to this report.

Endnotes

1 Microsoft, "Microsoft and OpenAI extend partnership," Press release, January 23, 2023, available at https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/; Amazon, "Amazon and Anthropic announce strategic collaboration to advance generative AI," Press release, September 25, 2023, available at https://www.aboutamazon.com/news/company-news/amazon-aws-anthropic-ai.

2 National Institute of Standards and Technology, "Application Programming Interface (API)," available at https://csrc.nist.gov/glossary/term/application_programming_interface (last accessed January 2024).

3 Google, "Responsible AI practices," available at https://ai.google/responsibility/responsible-ai-practices/ (last accessed January 2024); Microsoft, "Empowering responsible AI practices," available at https://www.microsoft.com/en-us/ai/responsible-ai (last accessed January 2024); OpenAI, "Developing safe & responsible AI," available at https://openai.com/safety (last accessed January 2024); Anthropic, "Core Views on AI Safety: When, Why, What, and How," March 8, 2023, available at https://www.anthropic.com/news/core-views-on-ai-safety; Amazon, "Transform responsible AI from theory into practice," available at https://aws.amazon.com/machine-learning/responsible-ai/ (last accessed January 2024); Meta, "Driven by our belief that AI should benefit everyone," available at https://ai.meta.com/responsible-ai/ (last accessed January 2024).

4 Microsoft, "Microsoft's AI Safety Policies," October 26, 2023, available at https://blogs.microsoft.com/on-the-issues/2023/10/26/microsofts-ai-safety-policies/.

5 Gartner, "Gartner Says More Than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026," Press release, October 11, 2023, available at https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026.

6 Besedo, "Building Trust and Safety: Why It Matters and How to Get It Right," November 7, 2023, available at https://besedo.com/knowledge-hub/blog/building-trust-and-safety-why-it-matters-and-how-to-get-it-right/.

7 Meta, "Facebook Community Standards," available at https://transparency.fb.com/policies/community-standards/ (last accessed January 2024).

8 YouTube, "Community Guidelines," available at https://www.youtube.com/howyoutubeworks/policies/community-guidelines/ (last accessed January 2024).

9 The White House, "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI," July 21, 2023, available at https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

10 Executive Office of the President, "Executive Order 14110: Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Press release, October 30, 2023, available at https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/; National Institute of Standards and Technology, "AI Risk Management Framework: Second Draft," August 18, 2022, available at https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf.

11 Executive Office of the President, "Executive Order 14110: Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."

12 Federal Trade Commission, "A Brief Overview of the Federal Trade Commission's Investigative, Law Enforcement, and Rulemaking Authority," available at https://www.ftc.gov/about-ftc/mission/enforcementauthority#:~:text=Section%206(b) %20empowers%20the,%2C%20and%20individuals.%5C"%2015%20U.S.C (last accessed January 2024).

13 Federal Trade Commission, "FTC Imposes $5 Billion Penalty and Sweeping New Privacy Restrictions on Facebook," Press release, July 24, 2019, available at https://www.ftc.gov/news-events/news/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions-facebook; *Federal Trade Commission v. Facebook*, complaint for civil penalties, injunction, and other relief, U.S. District Court for the District of Columbia, No. 19-cv-2184 (July 24, 2019), available at https://www.ftc.gov/system/files/documents/cases/182_3109_facebook_complaint_filed_7-24-19.pdf; *Federal Trade Commission v. Facebook*, stipulated order for civil penalty, monetary judgment, and injunctive relief, U.S. District Court for the District of Columbia, No. 19-cv-2184 (July 24, 2019), available at https://www.ftc.gov/system/files/documents/cases/182_3109_facebook_order_filed_7-24-19.pdf.

14 Megan Shahi and Adam Conner, "Priorities for a National AI Strategy," Center for American Progress, August 10, 2023, available at https://www.americanprogress.org/article/priorities-for-a-national-ai-strategy/.

15 Will Henshall, "4 Charts That Show Why AI Progress Is Unlikely to Slow Down," *Time*, August 2, 2023, available at https://time.com/6300942/ai-progress-charts/.

16 Executive Office of the President, "Executive Order 14110: Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."

17 Aditi Ganguly, "OpenAI's Meteoric Rise: $1 Billion In Annual Revenue On The Horizon," Yahoo Finance, September 21, 2023, available at https://finance.yahoo.com/news/openais-meteoric-rise-1-billion-173545014.html.

18 Jay Peters, "Google is bringing its AI-powered search to more than 120 new countries and territories," The Verge, November 8, 2023, available at https://www.theverge.com/2023/11/8/23951134/google-search -generative-experience-sge-expansion-120-countries-territories.

19 Kevin Scott, "Microsoft teams up with OpenAI to exclusively license GPT-3 language model," Microsoft, September 22, 2020, available at https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/.

20 StackCommerce, "Over 1 billion people worldwide use a MS Office product or service," *Financial Post*, April 10, 2021, available at https://financialpost.com/personal-finance/business-essentials/over-1-billion-people-worldwide-use-a-ms-office-product-or-service.

21 Jordan Novet, "Microsoft Office will now use AI to make meetings less painful," CNBC, November 15, 2023, available at https://www.cnbc.com/2023/11/15/microsoft-announces-new-copilot-features-in-outlook -powerpoint-teams.html.

22 Meta, "What's New Across Our AI Experiences," Press release, December 6, 2023, available at https://about.fb.com/news/2023/12/meta-ai-updates/.

23 Andrew R. Chow, "How ChatGPT Managed to Grow Faster Than TikTok or Instagram," *Time*, February 8, 2023, available at https://time.com/6253615/chatgpt-fastest-growing/.

24 The White House, "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI."

25 BSA | The Software Alliance, "AI Developers and Deployers: An Important Distinction" (Washington: 2023), available at https://www.bsa.org/policy-filings/ai-developers-and-deployers-an-important -distinction.

26 Ibid.

27 OpenAI, "Introducing ChatGPT," November 30, 2022, available at https://openai.com/blog/chatgpt; Anthropic, "Talk to Claude," available at https://claude.ai/login (last accessed January 2024); Google, "Bard," available at https://bard.google.com/ (last accessed January 2024); Microsoft, "Copilot," available at https://copilot.microsoft.com/ (last accessed January 2024); Meta, "Introducing New AI Experiences Across Our Family of Apps and Devices," September 27, 2023, available at https://about.fb.com/news/2023/09/introducing-ai-powered-assistants -characters-and-creative-tools/.

28 National Institute of Standards and Technology, "Application Programming Interface (API)."

29 European Parliament, "Artificial Intelligence Act," June 14, 2023, available at https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.

30 Alex Akimov, "It's critical to regulate AI within the multi-trillion-dollar API economy," TechCrunch, December 22, 2023, available at https://techcrunch.com/2023/12/22/its-critical-to-regulate-ai-within-the-multi-trillion-api-economy/.

31 Meta, "Llama 2 Version Release Date," July 18, 2023, available at https://ai.meta.com/llama/license/.

32 OpenAI, "Developer quickstart," available at https://platform.openai.com/docs/quickstart?context=python (last accessed January 2024).

33 OpenAI, "Introducing ChatGPT"; Anthropic, "Talk to Claude"; Google, "Bard"; Microsoft, "Copilot"; Meta, "Introducing New AI Experiences Across Our Family of Apps and Devices."

34 Reuters, "OpenAI plans major updates to lure developers with lower costs, Reuters sources say," CNBC, October 12, 2023, available at https://www.cnbc.com/2023/10/12/openai-plans-major-updates-to-lure- developers-with-lower-costs-reuters.html.

35 Greg Brockman and others, "Introducing ChatGPT and Whisper APIs," OpenAI, March 1, 2023, available at https://openai.com/blog/introducing-chatgpt-and-whisper-apis.

36 This is a slightly confusing example, as the LLM OpenAI developed is named ChatGPT and OpenAI's first-party deployment of ChatGPT is also known as ChatGPT. See Ibid.; Alex Heath, "Snapchat is releasing its own AI chatbot powered by ChatGPT," The Verge, February 27, 2023, available at https://www.theverge.com/2023/2/27/23614959/snapchat-my-ai-chatbot -chatgpt-openai-plus-subscription.

37 George Lawton, "Attributes of open vs. closed AI explained," TechTarget, August 25, 2023, available at https://www.techtarget.com/searchenterpriseai/feature/Attributes-of-open-vs-closed-AI-explained.

38 Meta, "Introducing Llama 2," available at https://ai.meta.com/llama/ (last accessed January 2024).

39 National Institute of Standards and Technology, "developer," available at https://csrc.nist.gov/glossary/term/developer (last accessed January 2024).

40 Sundar Pichai and Demis Hassabis, "Introducing Gemini: our largest and most capable AI model," Press release, December 6, 2023, available at https://blog.google/technology/ai/google-gemini-ai/.

41 National Institute of Standards and Technology, "Appendix A: Descriptions of AI Actor Tasks," available at https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Appendices/Appendix_A (last accessed January 2024).

42 Pichai and Hassabis, "Introducing Gemini: our largest and most capable AI model."

43 National Institute of Standards and Technology, "Appendix A: Descriptions of AI Actor Tasks."

44 Snapchat, "What is My AI on Snapchat, and how do I use it?", available at https://help.snapchat.com/hc/en-gb/articles/13266788358932-What-is-My-AI-on-Snapchat-and-how-do-I-use-it- (last accessed January 2024).

45 Meta, "Introducing Llama 2."

46 BigScience, "Introducing The World's Largest Open Multilingual Language Model: BLOOM," available at https://bigscience.huggingface.co/blog/bloom (last accessed January 2024).

47 Meta, "Llama 2: Responsible Use Guide" (Menlo Park, CA: 2023), available at https://ai.meta.com/static-resource/responsible-use-guide/; OpenAI, "GPT-4 System Card" (San Francisco: 2023), available at https://cdn.openai.com/papers/gpt-4-system-card.pdf.

48 OpenAI, "GPT-4 System Card."

49 Ibid.

50 James Manyika and Sissie Hsiao, "An overview of Bard: an early experiment with generative AI," October 19, 2023, available at https://ai.google/static/documents/google-about-bard.pdf.

51 Meta, "Llama 2: Responsible Use Guide."

52 Joshua J., "How should I report a GPT?", OpenAI, available at https://help.openai.com/en/articles/8554982-how-should-i-report-a-gpt (last accessed January 2024).

53 Besedo, "Building Trust and Safety: Why It Matters and How to Get It Right."

54 Amit Dar, "Why Generative AI Is The Next Frontier in Trust & Safety," ActiveFence, December 21, 2022, available at https://www.activefence.com/blog/generative-ai/.

55 Soumik Majumder, "Top Generative AI Industry Applications: An In-Depth Look," Turing, available at https://www.turing.com/resources/generative-ai-applications#7.-chatbot-performance-improvement (last accessed January 2024).

56 Gartner, "Gartner Experts Answer the Top Generative AI Questions for Your Enterprise," available at https://www.gartner.com/en/topics/generative-ai (last accessed January 2024).

57 Miles Brundage and others, "Lessons learned on language model safety and misuse," OpenAI, March 3, 2022, available at https://openai.com/research/language-model-safety-and-misuse#misuse.

58 OpenAI, "Usage policies," available at https://openai.com/policies/usage-policies (last accessed January 2024); Microsoft, "For Online Services," available at https://www.microsoft.com/licensing/ terms/product/ForOnlineServices/all (last accessed January 2024); Microsoft, "Code of conduct for Azure OpenAI Service," December 18, 2023, available at https://learn.microsoft.com/en-us/legal/cognitive -services/openai/code-of-conduct#responsible-ai-mitigation-requirements; Meta, "Llama Use Policy," available at https://ai.meta.com/llama/use-policy/ (last accessed January 2024).

59 OpenAI, "Usage policies."

60 Meta, "Llama Use Policy."

61 OpenAI, "Usage policies"; Microsoft, "For Online Services"; Microsoft, "Code of conduct for Azure OpenAI Service"; and Meta, "Llama Use Policy."

62 OpenAI, "Usage policies."

63 Meta, "Transparency Center," available at https://transparency.fb.com/ (last accessed January 2024); YouTube, "Community Guidelines," available at https://www.youtube.com/howyoutubeworks/policies/ community-guidelines/#enforcing-community-guidelines (last accessed January 2024).

64 Anthropic, "Acceptable Use Policy," September 15, 2023, available at https://console.anthropic.com/legal/aup.

65 Anthropic, "Core Views on AI Safety: When, Why, What, and How," March 8, 2023, available at https://www.anthropic.com/news/core-views-on-ai-safety.

66 Michael Schade, "ChatGPT iOS app - FAQ," available at https://help.openai.com/en/articles/7885016-chatgpt-ios-app-faq https://help.openai.com/en/articles/7885016-chatgpt-ios-app-faq (last accessed January 2024).

67 Anthropic, "How do I report harmful or illegal content?," available at https://support.anthropic.com/en/articles/7996906-how-do-i-report-harmful-or-illegal-content (last accessed January 2024).

68 OpenAI, "Introducing the GPT Store."

69 Gartner, "Gartner Says More Than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026."

70 Microsoft, "Content filtering," January 22, 2024, available at https://learn.microsoft.com/en-us/azure /ai-services/openai/concepts/content-filter?tabs=warning%2Cpython.

71 OpenAI, "Usage policies."

72 Sam Meredith, "Facebook-Cambridge Analytica: A timeline of the data hijacking scandal," CNBC, April 18, 2018, available at https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html.

73 Samuel Stolton, "Apple Set to Be Hit by EU Antitrust Order in App Store Fight With Spotify," Bloomberg, December 13, 2023, available at https://www.bloomberg.com/news/articles/2023-12-13/apple-set-to-be- hit-by-eu-antitrust-order-in-app-store-fight-with-spotify.

74 Gartner, "Gartner Says More Than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026."

75 OpenAI, "Terms of use," available at https://openai.com/policies/terms-of-use (last accessed January 2024); Anthropic, "Terms of Service," available at https://console.anthropic.com/legal/terms (last accessed January 2024).

76 OpenAI, "Terms of use."

77 OpenAI, "Usage policies"; Anthropic, "Commercial Terms of Service," available at https://www-files.anthropic.com/production/images/Anthropic-Commercial-Terms-of-Service_Dec_2023.pdf?dm=1703004244 (last accessed January 2024); Microsoft, "Microsoft – Terms of Use," available at https://www.microsoft.com/en-us/legal/terms-ofuse#:~:text=You%20are%20solely%20 responsible %20for,use%20of%20the%20AI%20 services (last accessed January 2024); Google, "Generative AI Additional Terms of Service," available at https://policies.google.com/terms/generative-ai (last accessed January 2024).

78 Michelle Cheng, "AI girlfriend bots are already flooding OpenAI's GPT store," Quartz, January 11, 2024, available https://qz.com/ai-girlfriend-bots-are-already-flooding-openai-s-gpt-st-1851159131.

79 OpenAI, "Usage policies."

80 Anthropic, "Acceptable Use Policy."

81 Microsoft, "Submit Abuse Report (CERT)," available at https://msrc.microsoft.com/report/abuse? Threa tType=URL&IncidentType=Responsible%20AI (last accessed January 2024).

82 OpenAI, "Chat model feedback," available at https://openai.com/form/chat-model-feedback (last accessed January 2024).

83 Canva, "Report content," available at https://www.canva.com/help/report-content/ (last accessed January 2024).

84 OpenAI, "How OpenAI is approaching 2024 worldwide elections," January 15, 2024, available at https://openai.com/blog/how-openai-is-approaching-2024-worldwide-elections.

85 Joshua J., "How should I report a GPT?".

86 OpenAI, "Safety best practices," available at https://platform.openai.com/docs/guides/safety -best-practices (last accessed January 2024).

87 Microsoft, "Code of conduct for Azure OpenAI Service."

88 Microsoft, "Content filtering."

89 OpenAI, "Safety best practices."

90 Anthropic, "Privacy Policy," available at https://console.anthropic.com/legal/privacy (last accessed January 2024).

91 See "How does OpenAI handle data retention and monitoring for API usage?" FAQ in OpenAI, "Enterprise privacy at OpenAI," available at https://openai.com/enterprise-privacy (last accessed January 2024).

92 Microsoft, "Data, privacy, and security for Azure OpenAI Service," June 23, 2023, available at https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy.

93 jota1, "Flagged to be in violation of policies," OpenAI, September 2023, available at https://community.openai.com/t/flagged-to-be-in-violation-of-policies/358979/4.

94 OpenAI, "Service terms"; Anthropic, "Commercial Terms of Service"; Microsoft, "Microsoft – Terms of Use"; Google, "Generative AI Additional Terms of Service."

95 David Evan Harris, "How to Regulate Unsecured 'Open-Source' AI: No Exemptions," Tech Policy Press, December 4, 2023, available at https://www.techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/.

96 Spandana Singh and Leila Doty, "The Transparency Report Tracking Tool: How Internet Platforms Are Reporting on the Enforcement of Their Content Rules," *New America*, December 9, 2021, available at https://www.newamerica.org/oti/reports/transparency-report-tracking-tool/.

97 National Institute of Standards and Technology, "NIST Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence," Press release, January 26, 2023, available at https://www.nist.gov/news-events/news/2023/01/nist-risk-management-framework-aims-improve-trustworthiness-artificial; National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" (Washington: U.S. Department of Commerce, 2023) available at https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

98 Executive Office of the President, "Executive Order 14110: Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."

99 National Institute of Standards and Technology, "Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)," *Federal Register* 88 (244) (2023): 88368–88370, available at https://www.federalregister.gov/documents/2023/12/21/2023-28232/request-for-information-rfi-related-to-nists-assignments-under-sections-41-45-and-11-of-the.

100 The White House, "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI."

101 Maria Diaz, "Third-party AI tools are responsible for 55% of AI failures in business," ZDNET, September 26, 2023, available at https://www.zdnet.com/article/third-party-ai-tools-are-responsible-for-55-of-ai-failures-in-business/.

102 Christchurch Call, "Four new tech firms expand the Christchurch Call," Press release, November 10, 2023, available at https://www.christchurchcall.com/media-and-resources/news-and-updates/four-new-tech-firms-join-the-christchurch-call.

103 Christchurch Call, "About: Christchurch Call text," available at https://www.christchurchcall.com/about/christchurch-call-text (last accessed January 2024).

104 National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)"; The White House, "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI"; Christchurch Call, "Four new tech firms expand the Christchurch Call"; Ibid.

105 OpenAI, "How OpenAI is approaching 2024 worldwide elections."

106 Ryan Heath, "AI's next battle: open or closed," June 26, 2023, available at https://www.axios.com/2023/06/26/ais-next-battle-open-closed-chatgpt; Lawton, "Attributes of open vs. closed AI explained."

107 Ibid.

108 Meta, "Introducing Llama 2."

109 Lawton, "Attributes of open vs. closed AI explained."

110 Dan Milmo, "Paedophiles using open source AI to create child sexual abuse content, says watchdog," *The Guardian*, September 13, 2023, available at https://www.theguardian.com/society/2023/sep/12/paedophiles-using-open-source-ai-to-create-child-sexual-abuse-content-says-watchdog; Heath, "AI's next battle: open or closed."

111 Heath, "AI's next battle: open or closed"; Amnesty International, "Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya" (London: 2022) available at https://www.amnesty.org/en/documents/ASA16/5933/2022/en/; Nicholas Confessore, "Cambridge Analytica and Facebook: The Scandal and the Fallout So Far," *The New York Times*, April 4, 2018, available at https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html.

112 Meta, "Welcome to Purple Llama," available at https://ai.meta.com/llama/purple-llama/ (last accessed January 2024).

113 Meta, "Llama Use Policy."

114 Elizabeth Seger and others, "Open-Sourcing Highly Capable Foundation Models" (Oxford: Centre for the Governance of AI, 2023), available at https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf.

115 OpenAI, "Usage policies"; Microsoft, "For Online Services"; Microsoft, "Code of conduct for Azure OpenAI Service"; Anthropic, "Acceptable Use Policy"; and Meta, "Llama Use Policy."

116 Ibid.; Anthropic, "Privacy Policy"; OpenAI, "Enterprise privacy at OpenAI."

117 Kevin Roose, "A.I. Poses 'Risk of Extinction,' Industry Leaders Warn," The New York Times, May 30, 2023, available at https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html.

118 Mary K. Pratt, "AI accountability: Who's responsible when AI goes wrong," TechTarget, August 19, 2021, available at https://www.techtarget.com/searchenterpriseai/feature/AI-accountability-Whos-responsible-when-AI-goes-wrong.

119 Ibid.

120 Katherine Haan, "24 Top AI Statistics And Trends In 2024," *Forbes*, April 25, 2023, available at https://www.forbes.com/advisor/business/ai-statistics/.

121 OpenAI, "Moderation," available at https://platform.openai.com/docs/guides/moderation (last accessed January 2024).

122 Microsoft, "Abuse Monitoring," July 18, 2023, available at https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/abuse-monitoring.

123 European Parliament, "Artificial Intelligence Act."

124 Executive Office of the President, "Executive Order 14110: Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence"; National Institute of Standards and Technology, "AI Risk Management Framework: Second Draft"; National Institute of Standards and Technology, "Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)."

125 National Institute of Standards and Technology, "Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)."

126 National Institute of Standards and Technology, "U.S. Artificial Intelligence Safety Institute" available at https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute (last accessed January 2024).

127 The White House, "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI."

128 Federal Trade Commission, "FTC Imposes $5 Billion Penalty and Sweeping New Privacy Restrictions on Facebook," Press release, July 24, 2019, available at https://www.ftc.gov/news-events/news/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions-facebook; Shana M. Broussard and Ellen L. Weintraub, "In the Matter of Cambridge Analytica LLC, et al.: Statement of Reasons of Chair Shana M. Broussard and Commissioner Ellen L. Weintraub," Federal Election Commission, November 4, 2021, available at https://www.fec.gov/files/legal/murs/7351/7351_99.pdf.

129 Lawyers' Committee for Civil Rights Under Law, "Online Civil Rights Act" (Washington: 2023), available at https://www.lawyerscommittee.org/online-civil-rights-act/.

130 Lawyers' Committee for Civil Rights Under Law, "Online Civil Rights Act: Highlights and Summary of Key Provisions" (Washington: 2023), available https://www.lawyerscommittee.org/wp-content/uploads/2023/12/LCCRUL_LC-Model-AI-Alg-Leg_summary.pdf.

131 Lilian Weng, Vik Goel, and Andrea Vallone, "Using GPT-4 for content moderation," August 15, 2023, available at https://openai.com/blog/using-gpt-4-for-content-moderation.

132 Snapchat, "Transparency Report," available at https://values.snap.com/privacy/transparency (last accessed January 2024); Google, "Google Transparency Report," available at https://transparencyreport.google.com/?hl=en (last accessed January 2024); Discord, "Transparency Reports," available at https://discord.com/safety-transparency-reports/2023-q1 (last accessed January 2024).

133 Perplexity, "Getting Started," available at https://blog.perplexity.ai/getting-started (last accessed January 2024).

americanprogress.org

1333 H Street, NW, 10th Floor, Washington, DC 20005 • Tel: 202-682-1611 • Fax: 202-682-1867