Acceptable Use Policies for Foundation Models

Kevin Klyman[1,2]

[1]Center for Research on Foundation Models, Stanford University
[2]Belfer Center for Science and International Affairs, Harvard University
kklyman@stanford.edu

Abstract

As foundation models have accumulated hundreds of millions of users, developers have begun to take steps to prevent harmful types of uses. One salient intervention that foundation model developers adopt is *acceptable use policies*—legally binding policies that prohibit users from using a model for specific purposes. This paper identifies acceptable use policies from 30 foundation model developers, analyzes the use restrictions they contain, and argues that acceptable use policies are an important lens for understanding the regulation of foundation models. Taken together, developers' acceptable use policies include 127 distinct use restrictions; the wide variety in the number and type of use restrictions may create fragmentation across the AI supply chain. Companies also employ acceptable use policies to prevent competitors or specific industries from making use of their models. Developers alone decide what constitutes acceptable use, and rarely provide transparency about how they enforce their policies. In practice, acceptable use policies are difficult to enforce, and scrupulous enforcement can act as a barrier to researcher access and limit beneficial uses of foundation models. Acceptable use policies for foundation models are an early example of self-regulation that have a significant impact on the market for foundation models and the AI ecosystem.

Introduction

Policymakers hoping to regulate generative AI have focused on preventing specific objectionable uses of AI systems, such as the creation of bioweapons (Mouton, Lucas, and Guest 2024), deepfakes (Chandra et al. 2024), and child sexual abuse material (Thorn 2024). Effectively blocking these uses can be difficult in the case of foundation models—large AI models trained on broad data that can be adapted to a wide range of downstream tasks—as they are general-purpose technologies that in principle can be used to generate any type of content (Bommasani et al. 2022). Yet developers of foundation models have been proactive, adopting policies as part of their terms of service or model licenses that prohibit potentially dangerous uses of the technology.

Foundation model developers have taken several approaches to adopting legally binding use restrictions. McDuff et al. (2024) find that developers of open-weight foundation models increasingly distribute models with licenses that include a standardized set of behavioral use restrictions. Developers of closed-weight models also restrict how users can make use of their models via terms of service agreements that prohibit generating specific categories of content (Bommasani et al. 2023b). Developers often refer to policies that include legally binding use restrictions on foundation models as acceptable use policies (AUPs), as they determine the domains of use that are acceptable and prohibited.

This paper collates and analyzes the acceptable use policies of 30 foundation model developers in order to assess their impact. It addresses the following question: what do acceptable use policies reveal about the ways that foundation model developers seek to regulate end-user behavior, and how do they impact the foundation model ecosystem?

The paper proceeds as follows: the background section defines acceptable use policies for foundation models, comparing them to similar policies for other technologies and to documents like model cards. The methodology section lays out the methods used to identify acceptable use policies and analyze their content. The analysis section describes the differences between developers' policies in terms of prohibited content and restrictions on types of end use. The enforcement section outlines difficulties in policy enforcement and potential downsides from strict enforcement. The discussion section highlights developers' decision-making power, how gaps in use restrictions may facilitate misuse, and how acceptable use policies shape the foundation model market.

Background

What Is an Acceptable Use Policy?

Acceptable use policies are common across digital technologies (O'Byrne 2019). Providers of public access computers (Robinson and McMenemy 2020), websites (Stewart 2000; Weidman and Grossklags 2019), and digital platforms (Siau, Nah, and Teng 2002; Pater et al. 2016) have long adopted acceptable use policies that articulate how their terms of service restrict what users can and cannot do with their products and services. While enforcement of these policies is uneven, restrictions on specific uses of digital technologies are widespread (Doherty, Anastasakis, and Fulford 2011).

The acceptable use policies of social media companies (CIP et al. 2021), cloud service providers (Gregorio et al. 2019), and content delivery networks (Looney 2023) have

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

received scrutiny as they constrain the behavior of billions of users. Acceptable use policies adopted by employers, which limit employees' use of company-provided technologies (Li, Zhang, and Sarathy 2010), schools, which limit students' use of the internet (Ahn, Bivona, and DiScala 2011), and public libraries, which limit the public's use of public access computers (McMenemy 2019), have come into focus as issues related to enforcement arise.

In the context of foundation model development, an acceptable use policy is a policy from a developer that determines how a foundation model can or cannot be used. Acceptable use policies restrict the use of foundation models by detailing the types of content users are prohibited from generating as well as domains of prohibited use.[1] Developers make these restrictions legally binding by including acceptable use policies in terms of service agreements or in copyright licenses for their foundation models.

Acceptable use policies typically aim to prevent users from using a foundation model to generate content that may violate the law or otherwise cause harm.[2] They accomplish this by listing specific subcategories of violative content and authorizing model developers to punish users who generate such content by, for example, limiting the number of queries users can issue or banning a user's account.

Acceptable use policies relate to how foundation models are built in important ways. For example, developers filter training data to remove content relevant to requests that would violate their acceptable use policies. OpenAI's GPT-4 technical report states: "We reduced the prevalence of certain kinds of content that violate our usage policies (such as inappropriate erotic content) in our pre-training dataset, and fine-tuned the model to refuse certain instructions such as direct requests for illicit advice" (OpenAI et al. 2024).

In addition, some developers state that the purpose of reinforcement learning from human feedback (RLHF) and safety fine-tuning is to make their foundation models less likely to generate outputs that would violate their acceptable use policies. Anthropic's model card for Claude 3 notes "We developed refusals evaluations to help test the helpfulness aspect of Claude models, measuring where the model unhelpfully refuses to answer a harmless prompt, i.e. where it incorrectly categorizes a prompt as unsafe (violating our AUP) and therefore refuses to answer" (Anthropic 2024).

How Do AUPs Differ From Other Documents?

Acceptable use policies are not the only way developers restrict use of their models. Other policy-related mechanisms that developers implement to restrict model use include:

- *Model Cards*: Model cards, which are published alongside machine learning models, provide essential information about models such as their intended uses and out-of-scope uses (Mitchell et al. 2019). However, model cards are not enforceable contracts, and they are not generally referenced in model licenses or developers' terms of service; as a result, out-of-scope uses do not rise to the same level as prohibited uses in an acceptable use policy.
- *Model Behavior Policies*: Model behavior policies determine what a model can or cannot do (OpenAI 2024; Google 2024; Bai et al. 2022). While acceptable use policies apply to user behavior, model behavior policies apply to the behavior of the model itself (Bommasani et al. 2023b). A model behavior policy is one way of embedding an acceptable use policy into a model; methods for imposing a model behavior policy include using RLHF to cause the model to be more likely to refuse violative prompts or employing a safety classifier at inference time to filter violative model outputs (Inan et al. 2023). Model behavior policies are generally broader than acceptable use policies; for instance, many developers fine-tune their models to produce more polite responses, though they do not block users from generating impolite responses.
- *Third party contracts*: Foundation model developers frequently partner with other firms to disseminate foundation models (Cen et al. 2023). These include cloud service providers (e.g., AWS, Azure, GCP), platform providers (e.g., Scale AI, Nvidia), database providers (e.g., Salesforce, Oracle), and model distributors (e.g., Together, Quora) (Srikumar et al. 2024). Custom contracts with third party providers of a developer's foundation models often include use restrictions, but the extent to which companies' acceptable use policies are altered via these partnership agreements is unclear.

Norms and Laws on Acceptable Use Policies

Although generative AI is a nascent industry, norms have begun to emerge around use restrictions for foundation models. Cohere, OpenAI, and AI21 Labs (2022) wrote in their "Best practices for deploying language models" that organizations should "[p]ublish usage guidelines and terms of use of LLMs in a way that prohibits material harm...such as through spam, fraud, or astroturfing." Developers of open-weight foundation models often adopt the same acceptable use policies by reusing the same model licenses. For example, more than 3,000 models on Hugging Face use Meta's Llama 2 license (McDuff et al. 2024).

Governments have taken an interest in acceptable use policies, which are a salient effort by foundation model developers to self-regulate (Ferretti 2022). Annexes IXa and IXb of the EU AI Act require that all providers of general-purpose AI models disclose the "acceptable use policies [that are] applicable" to both the EU's AI Office and other firms that integrate the general-purpose AI model into their own AI systems (EU 2024; Hacker 2023). China's Interim Measures for the Management of Generative AI Services, which were adopted in July 2023, go a step further by requiring that providers of generative AI services act to prevent users from "using generative AI services to engage in illegal activities. . . including [by issuing] warnings, limiting

[1] This differs from the case of non-generative technologies, where restrictions focus on a user's input not a system's output.

[2] Acceptable use policies for 3D printers, another generative, general-purpose technology with the potential to cause real-world harm (Beyer 2014), are among the best analogue for the case of foundation models. Public libraries have adopted AUPs that prohibit 3D printing of ghost guns, sex toys, and swastikas, for instance (Jones 2015; Minow et al. 2016).

functions, and suspending or concluding the provision of services" (CAC 2023; Zhang 2024). And the US Voluntary AI Commitments require firms to publicly report "domains of appropriate and inappropriate use" as well as limitations of the model that affect these domains (White House 2023).

Methodology

Search Protocol for Acceptable Use Policies

Table 1 details 30 foundation model developers' acceptable use policies. Developers use different policy documents to limit model use, including: a standalone acceptable use policy for all their foundation models (e.g., Google, Stability AI), use restrictions included in a general model license (e.g., AI2), use restrictions included in a custom model license (e.g., BigScience, Meta), or provisions in terms of service agreements that apply to all services including foundation models (e.g., Midjourney, Perplexity, Eleven Labs).

The following protocol was used to identify acceptable use policies across these different types of documents:

1. Compile a list of foundation model developers using the data provided by Bommasani et al. (2023c).
2. For each developer, check the terms of service (TOS) on its website. If the TOS include an AUP with content restrictions that plausibly cover the developer's foundation models, take that part of the TOS as the AUP.
3. For each remaining developer, check the license for its "flagship foundation model"; if it includes behavioral use restrictions, take that part of the license as the AUP.[3]
4. For each remaining developer, if the TOS or license reference a separate document with behavioral use restrictions such that the restrictions are binding, take the relevant portion of that document as the AUP.

Coding of Prohibited Use Categories in AUPs

Qualitative content analysis was used for this paper's coding of prohibited use categories in developers' acceptable use policies (Mayring 2015). This was done inductively (Elo and Kyngäs 2008), with categories drawn directly from acceptable use policies, and was inspired by prior work related to AI ethics guidelines (Fjeld et al. 2020), privacy policies (Alfawzan et al. 2022), content moderation guidelines (CIP et al. 2021), benchmarks (Wang et al. 2024a), and Responsible AI Licenses (McDuff et al. 2024).

The following process was used to code the prohibited use categories included in developers' acceptable use policies:

- For each acceptable use policy, each line of the policy was analyzed. For each line, the distinct prohibited use categories included were added to a list of prohibited uses across every developers' acceptable use policy. Distinct prohibited use categories do not include different types of actions related to the same prohibited use category (e.g., "generating, promoting, or further distributing spam" was coded as "spam") or categories with substantial overlap that do not use distinct phrasing.
- Using the list of prohibited use categories across all AUPs, each line of each acceptable use policy was considered again to ensure the prohibited use categories therein are coded correctly. A prohibited use category should receive a specific coding only if it uses near-identical language to that coding, and each prohibited use category in each policy receives only one coding.

This produced a list of 127 categories and a 30x127 matrix, visible at github.com/kklyman/aupsforfms, where columns show developers, rows show prohibited use categories, and cells are marked "1" if a developer's acceptable use policy explicitly references that prohibited use category and "0" otherwise. Section 4 analyzes the results of this coding.

This methodology satisfies three aims. First, it provides a systematic and comprehensive approach for capturing the prohibited use categories included in acceptable use policies. Second, it enables a granular analysis of acceptable use policies. Classifying prohibited use categories into higher-level groups is an illustrative exercise (see Figure 1), but acceptable use policies are legal documents with unique provisions that require close study (O'Byrne 2019). Third, it clarifies the risks from foundation models that developers themselves seek to mitigate. While many previous works have taxonomized the risks and harms stemming from foundation models (Weidinger et al. 2023; Shelby et al. 2023; Hoffmann and Frase 2023; Fergusson et al. 2023), this paper assesses how firms taxonomize risk on the basis of their own policies.

Analysis of Acceptable Use Policies

Developers With Acceptable Use Policies

Foundation model developers that have AUPs are heterogeneous along multiple axes, demonstrating broad adoption (see Table 1). In terms of model release, 12 of the developers openly release the model weights for their flagship model series, while 18 do not. These models have a variety of different output modalities, with 20 language models, 4 multimodal models, 3 image models, 2 video models and 1 audio model. The developers are headquartered around the world, with 19 based in the US and the others based in Canada, China, France, Germany, Israel, UAE, and the UK.

Prohibited Content in Acceptable Use Policies

Acceptable use policies commonly prohibit users from employing foundation models to generate content that is explicit (e.g., violence, pornography), fraudulent (scams, spam), abusive (harassment, hate speech), deceptive (disinformation, impersonation), or otherwise harmful (malware, privacy infringements).[4] Figure 1 shows the most common categories of content that are explicitly prohibited by developers' acceptable use policies: mis/disinformation (26 policies include explicit prohibitions), harassment/abuse (26), privacy (21), discrimination (21), and child harm/child sexual abuse material (21) were the most frequent, while categories like political content (9), medical advice (8), weapons

[3] A flagship foundation model is a developer's most salient and/or capable model, informed by its public documentation.

[4] Content-based restrictions generally apply only to user prompts that request that a model generate this type of content—models will classify the toxicity of this type of content if asked to do so, but it is against developers' policies to generate such content.

Developer	Title of Acceptable Use Policy, Section Including Use Restrictions	Flagship Model Series (Output Modality)	HQ	Open / Closed
01.ai	Community License Agreement v2.1, §2 License and License Restrictions	Yi (Text)	PRC	Open
Adept	Terms of Use, §1.1(d) Usage Restrictions	Fuyu (Multimodal)	USA	Open
Adobe	Generative AI User Guidelines	Firefly (Image)	USA	Closed
AI21	Usage Guidelines	Jurassic-2 (Text)	ISR	Closed
AI2	AI2 ImpACT License for Low-Risk Artifacts	OLMo (Text)	USA	Open
Aleph Alpha	Terms and Conditions, §4.8 Customer's Rights and Restrictions	Luminous (Text)	DEU	Closed†
Amazon	AWS Responsible AI Policy & Acceptable Use Policy*	Titan Text (Text)	USA	Closed
Anthropic	Acceptable Use Policy	Claude 3 (Text)	USA	Closed
Baidu	User Agreement, §4 Service Usage Specifications	ERNIE 4.0 (Text)	PRC	Closed
BigCode	BigCode Open RAIL-M v1 License, §A Use Restrictions	StarCoder 2 (Text)	N/A‡	Open
BigScience	BigScience RAIL License v1.0, §A Use Restrictions	BLOOM (Text)	N/A‡	Open
Character.AI	Terms of Service, Conditions of Use	Not Public (Text)	USA	Closed
Cohere	Usage Guidelines	Command (Text)	CAN	Closed
Databricks	Databricks Open Model Acceptable Use Policy	DBRX (Text)	USA	Open
DeepSeek	Terms of Use, §3 Service Management	DeepSeek (Text)	PRC	Open
Eleven Labs	Terms of Service, Prohibited Activities	Not Public** (Audio)	USA	Closed
Google	Generative AI Prohibited Use Policy	Gemini (Multimodal)	USA	Closed
Inflection	Terms of Service, Acceptable Use	Inflection-2.5 (Text)	USA	Closed
Meta	Acceptable Use Policy	Llama 2 (Text)	USA	Open
Midjourney	Terms of Service, §9 Community Guidelines	Midjourney v6 (Image)	USA	Closed
Mistral	Terms of Use, §8 Your obligations/§9 Our Obligations & Le Chat Terms of Service, §4.3 Chat Moderation Policy*	Mixtral (Text)	FRA	Open
OpenAI	Usage Policy	GPT-4 (Multimodal)	USA	Closed
Perplexity	Terms of Service, Acceptable Use	Not Public** (Text)	USA	Closed
Reka	Terms of Service, §3.2 Responsible Use	Yasa-1 (Multimodal)	USA	Closed
Runway	Terms of Service, §5 User Conduct	Not Public** (Video)	USA	Closed
Stability AI	Acceptable Use Policy	Stable Diffusion 3 (Image)	GBR	Open
TII	Acceptable Use Policy	Falcon 180B (Text)	UAE	Open
Together	Terms of Service, §2.4 Your Responsibilities	StripedHyena Nous (Text)	USA	Open
Twelve Labs	Terms of Service, §14 No Unlawful or Prohibited Use	Pegasus-1 (Video)	USA	Closed
Writer	Terms and Conditions, §4.3 Acceptable Use	Palmyra-1 (Text)	USA	Closed

Table 1: 30 Foundation Model Developers' AUPs. *Amazon and Mistral's TOS explicitly refer to two relevant documents, so both are considered. **These developers have not publicly disclosed details about their flagship foundation models. †Aleph Alpha provides model weights to customers on prem. ‡International research groups. (Last updated 4/18/24. See GitHub.)

(7), surveillance (7), and plagiarism (4) were less common.

Many developers' acceptable use policies have granular use restrictions, whereas others have broad restrictions without much elaboration. Figure 1 shows the number of prohibited use categories in each developers' acceptable use policy and distinguishes between open- and closed-weight developers (Kapoor et al. 2024). Among closed developers, the acceptable use policies of Anthropic (69 prohibited uses), Cohere (46), and OpenAI (46) explicitly reference the largest number of prohibited use categories, while the policies of smaller startups such as Reka (15), Writer (14), and Perplexity (12) have the fewest. Among open developers, the acceptable use policies of Stability AI (44), Meta (44), and Mistral (38) explicitly reference the largest number of prohibited use categories, while the AUPs of 01.ai (11), Together (7), and the Technology Innovation Institute (6) have the fewest. The average number of prohibited uses for closed developers is 20 (standard deviation of 15.1), while the average for open developers is 24.5 (standard deviation is 13.5).

There are several potential explanations for open developers having a larger number of prohibited use categories in their AUPs. Open foundation model developers often use Responsible AI Licenses that feature a sizable, standardized set of use restrictions (Contractor et al. 2022; Keller and Bonato 2023). Second, a greater number of closed foundation model developers have acceptable use policies (including smaller companies without large legal teams), whereas many other open developers have no acceptable use policy, introducing potential selection bias in computing the average. Third, unlike closed developers, open developers often cannot enforce their policies against individual users, so prohibiting a larger number of uses may come at less cost.

The strength of an acceptable use policy is not determined solely by the number of prohibited uses it lists. All 30 ac-

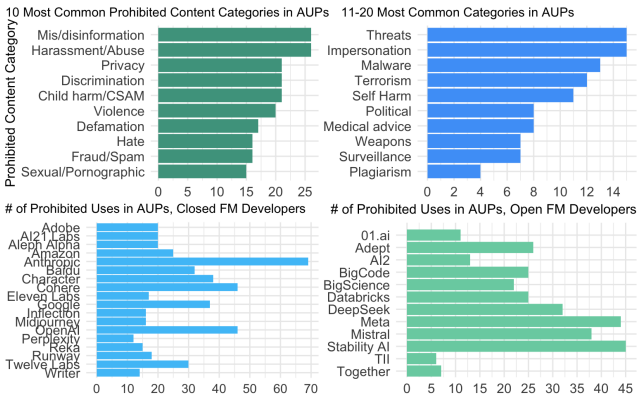

Figure 1: Common prohibited content categories and number of prohibited uses per developer. Top left: the 10 most common categories of content-related prohibited uses in developers' AUPs. Top right: the next 10 most common categories of content-related prohibited uses in developers' AUPs. (See the GitHub for details on grouping.) Bottom left: the number of explicitly prohibited uses in closed developers' AUPs (out of 127 categories). Bottom right: the number of explicitly prohibited uses in open developers' AUPs.

ceptable use policies prohibit users from generating content that violates the law, and the majority prohibit users from generating content that impedes the model developer's operations or is not accompanied by adequate disclosure that it is machine-generated. These catch-all prohibitions cover unenumerated risk categories, making acceptable use policies more malleable and comprehensive by linking them to laws and organizational procedures that may change. Over 40 of the 127 prohibited use categories relate to potentially illegal content (e.g., child sexual abuse material, defamation, discrimination against a protected class, drugs, fraud, hate speech, malware, prostitution, scams), reflecting that developers consider their models to be capable of generating illegal content and wish to reduce their liability for such risks.

Prohibitions on content that is not generally illegal show developers' priorities and highlight different approaches taken in their acceptable use policies. Political content, such as using foundation models for campaigning, lobbying, or otherwise influencing political processes, is explicitly prohibited by 9 startups—Anthropic, Character.AI, Cohere, Databricks, Midjourney, OpenAI, Perplexity, Stability AI, and Twelve Labs—whereas Big Tech companies like Amazon, Google, and Meta have no such prohibitions. Weapons-related content is explicitly prohibited by 7 developers: AI2, Anthropic, Amazon, Meta, Mistral, OpenAI, and Stability AI. Generating eating disorder-related content, such as pro-anorexia content, is explicitly prohibited by just 4 developers: Character.AI, Cohere, Meta, and Mistral. And while some open developers such as Adept, DeepSeek, and Together broadly prohibit some types of sexual content, others like Meta and Mistral prohibit only content related to prostitution or sexual violence. Foundation models have the potential to cause harm in these areas, yet major developers choose not to adopt legally binding restrictions on these uses of their models (Goldstein et al. 2024; Suchman 2020).

Other notable prohibited uses include:

- *Undermining the interests of the state*: Baidu and DeepSeek, two of three model developers headquartered in China, state in their acceptable use policies that users must not generate content "endangering national security, leaking state secrets, subverting state power, overthrowing the socialist system, and undermining national unity…damaging the honor and interests of the state…undermining the state's religious policy". 01.ai, the other Chinese developer, also includes a prohibition against "harming national security." These restrictions draw directly on China's rules on foundation model-generated content (Zeng and Klyman et al. 2024).
- *Password trafficking*: Eleven Labs, the only developer whose flagship model outputs audio, prohibits users from using its models to "trick or mislead us or other users, especially in an attempt to learn sensitive account information, for example user passwords." This may be intended to address concerns regarding the use of voice cloning for scams (Marchal et al. 2024; Barnett 2023).
- *Misinformation*: The extent to which developers restrict users' ability to generate and/or distribute inaccurate content varies widely. While some AUPs include wholesale bans on misinformation (e.g., AI21 Labs, Inflection), others have looser restrictions that apply only to verifiable disinformation with the intent to cause harm (e.g., TII). Mis/disinformation is the most frequently prohibited category of use—even more so than child sexual abuse material—indicating some developers may be more responsive to political and reputational risk than assessments of harm or legal liability (Pfefferkorn 2024).

Restrictions on Types of End Use

In addition to content-based restrictions, acceptable use policies for foundation models often restrict the types of activity that users can carry out. Acceptable use policies from 6 developers prohibit "model scraping" or training a model on their own model's outputs. Anthropic's Acceptable Use Policy bans use of "prompts and results to train an AI model (e.g., 'model scraping')"; Adept, Adobe, Meta, Perplexity, and Runway similarly prohibit the use of model outputs for training other foundation models. While 8 developers have no such explicit ban (BigCode, BigScience, Character.AI, Eleven Labs, Mistral, Stability AI, TII, Reka), the remaining 16 prohibit use of their models to build a competing service, which encompasses model scraping (Metz and Ford 2024).

Some developers prohibit distribution of AI-generated content at scale. AI21 Labs' Usage Guidelines state that "No content generated by AI21 Studio will be posted automatically (without human intervention) to any public website or platform where it may be viewed by an audience greater than 100 people." Four other developers (BigCode, BigScience, Cohere, and Databricks) prohibit using their models for automated posting online (Goldstein et al. 2023).

Many acceptable use policies prevent firms in certain industries from making use of foundation models. For example, weapons manufacturers would be in violation of a pol-

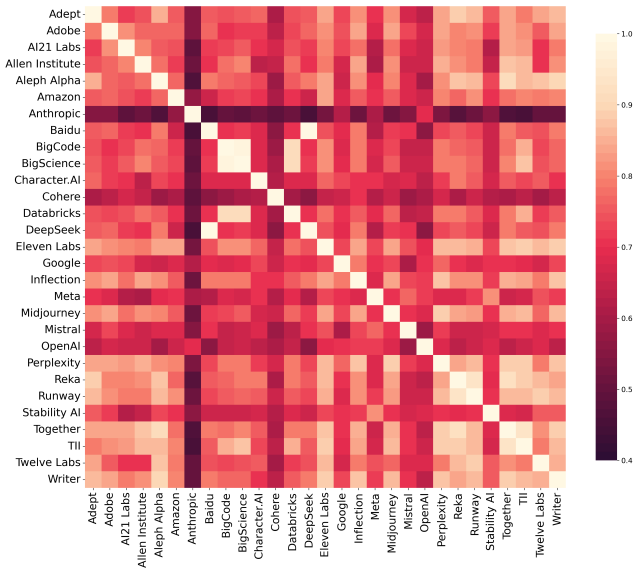

Figure 2: Developer correlations. The correlation between prohibited use categories for pairs of developers across all 127 categories. Correlation is measured using the simple matching coefficient (i.e. agreement rate), which is the fraction of all indicators for which both developers are assigned the same value (i.e. where both are assigned 1 as both of their AUPs prohibit the category, or both are assigned 0).

icy with weapons-related restrictions if they made use of the foundation model to produce weapons, though it is possible that the developer negotiates custom contracts with weapons manufacturers (Brenes and Hartung 2024). In January 2024, OpenAI reportedly changed its Usage Policies to facilitate partnerships with militaries, deleting a line that prohibited use related to "military and warfare" (Biddle 2024).

Acceptable use policies may also restrict the use of models in highly-regulated industries such as law, finance, and medicine. 8 of the 30 acceptable use policies include restrictions on medical advice, and Anthropic, Character.AI, Google, Meta, and OpenAI also have restrictions on legal and financial advice, which apply not only to lawyers, doctors, and financial advisers, but also to organizations that provide services in these fields.

AI2, Amazon, Anthropic, Google, and OpenAI also prohibit use of their models for certain types of surveillance. Google prohibits use of its models for "tracking or monitoring people without their consent" while AI2 singles out "military surveillance." This could prevent spyware companies and intelligence contractors respectively from making use of their foundation models (Feldstein 2019).

Correlations Between Developers' AUPs

Despite increased standardization across open developers (McDuff et al. 2024), AUPs remain inconsistent across foundation model developers. Figure 2 shows the correlation between developers' acceptable use policies based on which of the 127 prohibited use categories they include. BigCode, BigScience, and Databricks have highly similar policies (with a correlation of more than 0.9), as do Baidu and DeepSeek (two Chinese developers) and Reka, TII, and Together (developers with few prohibited use categories). Anthropic, Cohere, Google, Meta, Mistral, OpenAI, and Stability AI are among the developers with the policies that are least similar to others, in part because they have the largest number of prohibited uses; each have a correlation of 0.7 or less with 15 or more other developers. This may pose an issue for cloud providers that distribute models from many developers; Amazon Web Services, for example, distributes models from AI21 Labs, Amazon, Anthropic, Cohere, Meta, Mistral, and Stability AI, but Anthropic's acceptable use policy has a correlation of less than 0.6 with those each of these other developers, indicating AWS would need to enforce several substantially different policies.

Developers Without Acceptable Use Policies

There are tens of developers that do not have acceptable use policies for their foundation models—Table 2 provides 7 examples. There are a number of reasons why a developer may choose to release a model without an acceptable use policy. Some open developers do not use acceptable use policies because their models are intended for research only—use restrictions could deter safety research via adversarial red teaming (in the absence of a safe harbor for good faith researchers) (Basdevant et al. 2024; Longpre et al. 2024b). Other models intended for research may lack acceptable use policies on the basis that they present less severe risks of misuse, whether because they have less significant capabilities or fewer users. Non-commercial models such as these are frequently distributed using licenses without use restrictions like Apache 2.0 or Creative Commons Attribution-NonCommercial licenses (White et al. 2024; Longpre et al. 2024a). While a license may not include any use restrictions for noncommercial users, commercial users may have to agree to custom use restrictions in their contracts with the model developer, which are not public. This creates a potential information asymmetry where a developer and its clients know the domains of permitted use, while regulators and the public may be led to believe that model use is unrestricted (Hacker, Cordes, and Rochon 2024).

Foundation models available for commercial use may not include AUPs for several reasons. In some cases, developers offer a model "as is," stating that it is not intended for commercial use without further fine-tuning, mitigations, or use restrictions (e.g., Databricks' MPT-30B). Developers hoping to maximize uptake among commercial users may be less likely to adopt acceptable use policies because clients' legal teams may recommend using different models without such restrictions. Other developers release their models without complete documentation, whether because they intend to release an acceptable use policy at a later point, which could be part of staged release, or due to under-documentation in the rush to release a model (Solaiman 2023).

In any case, other restrictions may apply to foundation models without acceptable use policies. Alibaba Cloud restricts firms with over 100 million users from making use of Qwen-VL through its license, which also bans model scraping. Restrictions on who can use a foundation model may

have a significant effect on how it is used even in the absence of legally binding behavioral use restrictions.

Enforcement of Acceptable Use Policies

Barriers to Enforcement

Practical and Legal Barriers for Open Developers The enforceability of open foundation model developers' acceptable use policies is a major limitation on how effective they are at restricting risky uses. Unlike closed foundation model developers, whose models are distributed via their own products, services, or APIs (or those of another firm), developers of open foundation models distribute their models by sharing the weights online such that they can be downloaded, and models are often run locally (Solaiman 2023). As a result, open developers have few ways of monitoring downstream use of their models, making it difficult to enforce their policies where models are run locally or where hosted inference is provided by another organization.

If open foundation model developers were to attempt to enforce their acceptable use policies, many would face substantial legal barriers. Licenses for open-weight foundation models that include behavioral use restrictions are a type of copyright license, but it is unclear if machine learning models are copyrightable artifacts, calling into question the enforceability of such licenses (Henderson et al. 2023). Downing (2023) argues that even if Responsible AI Licenses for models do not trigger copyright issues, the use restrictions in these licenses are ineffective as licensees are not required to enforce them against downstream licensees and developers cannot sue downstream licensees for violations. Licenses for open-weight models also face issues related to interoperability, as use restrictions may not propagate to software that receives inputs from the model (Downing 2024).

Nevertheless, private sector licensees will likely comply with acceptable use policies of open-weight foundation models due to the legal risk associated with noncompliance. In cases where an open developer does not seek to enforce its acceptable use policy, the policy can still encourage responsible use (Pistilli et al. 2023). Most users are not bad actors and may adhere to a policy despite gaps in enforcement, as they have no interest in generating prohibited content.

Despite these challenges, many open foundation model developers attempt to restrict the use of their models to some degree. 12 of the developers that have acceptable use policies openly release their flagship model's weights, but do so using licenses or terms of service that prohibit certain uses. Although open foundation models are frequently referred to as "open-source" in popular media, truly open-source software or machine learning models cannot have use restrictions by definition (OSI 2024; Downing 2024).

Ecosystem Barriers Another issue in gauging the enforcement of acceptable use policies is the way in which they propagate across the AI ecosystem. In addition to developers, cloud service providers (e.g., AWS, Azure, GCP) and other digital platforms (e.g., Salesforce, Scale AI) act as deployers of foundation models that were not developed in-house. Deployers' platforms have their own acceptable use policies that do not align perfectly with external developers' acceptable use policies, and it is not clear that a deployer would have adequate expertise to restrict the uses of a foundation model in accordance with an acceptable use policy that is more stringent than that of the deployer (Gregorio et al. 2019). In particular, deployers would need to build infrastructure to support enforcement of the distinct acceptable use policies for each of the foundation models they distribute. While there are a variety of publicly available models and tools that deployers might leverage to enforce developers' acceptable use policies (e.g., by filtering specific categories of prompts and responses), there is little evidence deployers have done so.

As an alternative, a deployer may attempt to devise (and enforce) its own acceptable use policy that encompasses those of each of its developer partners. However, the large variation in prohibited use categories among different developers' acceptable use policies makes such an exercise difficult, and may require that for each category the deployer apply the most restrictive of its partners' acceptable use policies to every model. Gorwa and Veale (2024) find that model marketplaces such as Hugging Face and GitHub have struggled to enforce their own acceptable use policies in light of the challenge of moderating the distribution of thousands of machine learning models, each of which may come with its own use restrictions (Hugging Face 2023; GitHub 2024).

These challenges are made more stark by the ease with which users can circumvent technical measures used to enforce acceptable use policies. Zeng et al. (2024) show that including uncommon dialects and appeals to authority in prompts can cause a foundation model to violate its developer's acceptable use policy despite safety filters in APIs. In addition, Qi et al. (2023) find that fine-tuning foundation models via an API can remove safety measures like instruction tuning and RLHF such that models will more readily violate their developer's acceptable use policy. Other researchers have found many vulnerabilities that allow users to nullify measures intended to promote adherence to acceptable use policies, such as adversarial prompts (Maus et al. 2023), jailbreaks (Wei, Haghtalab, and Steinhardt 2023; Zou et al. 2023; Shah et al. 2023), and other methods for fine-tuning away safety measures via APIs (Yang et al. 2023; Zhan et al. 2023). These vulnerabilities show that closed developers are likely unable to enforce their acceptable use policies in many cases (Henderson et al. 2024).

Misallocating Responsibility to Users Acceptable use policies are a means of shifting responsibility (and liability) for risky uses of a technology from the developer, deployer, or distributor of that technology to the user (Doherty, Anastasakis, and Fulford 2011; Weidman and Grossklags 2019). Acceptable use policies may be effective in limiting the behavior of corporate users, which are legally risk-averse, but are unlikely to fundamentally alter the behavior of the average individual user (Villa 2022).

Developers' approach to indemnification crystallizes the issue. Meta's Llama licenses, for example, hold users responsible for any direct or downstream use of the model, stating "[y]ou will indemnify and hold harmless Meta from and against any claim by any third party arising out of or

Developer	Model	Intended Use	Model Assets Released	License
Alibaba Cloud	Qwen-VL	"Researchers and developers are free to use the codes and model weights"	Code, weights	Tongyi Qianwen License
EleutherAI	GPT-NeoX 20B	"Developed primarily for research purposes… not intended for deployment as-is."	Data, code, weights	Apache 2.0
! Meta	MusicGen-Large	"model should not be used on downstream applications without further risk evaluation and mitigation."	Data, code, weights	CC-BY-NC 4.0
Microsoft	Phi-2	"Direct adoption for production tasks without evaluation is out of scope of this project."	Weights	MIT
Mistral	Mixtral-8x7B	N/A	Code, weights	Apache 2.0
Databricks	MPT-30B	"Not intended for deployment without finetuning."	Code, weights	Apache 2.0
xAI	Grok-1	"intended to be used... [for] question answering, information retrieval, creative writing and coding"	Weights	Apache 2.0

Table 2: Foundation Model Developers Without AUPs. Information on developers without acceptable use policies, including the name of the developer, the name of the model where no acceptable use policy has been applied, the intended use of that model (from the model card but for Alibaba Cloud), model assets released, and the license under which the model is distributed.

related to your use or distribution of the Llama Materials."

Social media companies' content policies also shift responsibility for toxic content from platforms that algorithmically amplifies such content to individual users that post it (Keller 2021). The same can be said of AI ethics guidelines, which often provide guidance to users regarding how to ethically use a company's AI systems rather than describing the tangible steps a company will take to prioritize ethics above other aims (Chi, Lurie, and Mulligan 2021; Fjeld et al. 2020). Similarly, developers employ acceptable use policies to eschew responsibility for downstream impacts of the foundation models they choose to build and deploy.

Acceptable use policies often impose obligations on users that they are ill-equipped to uphold. Setting aside issues of digital literacy (Ng et al. 2021), the user is often not the right party to be responsible for ensuring that a foundation model is not generating violative outputs (Cooper et al. 2022). For instance, holding users responsible for generating self-harm related content may be viable for users that maliciously seek to spread such content online, but not for vulnerable users seeking to harm themselves and who turn to a model for aid.

One solution that developers implement is increasing surveillance of their users to monitor dangerous prompts and responses. Robinson (2019) argues surveillance is a fundamental feature of acceptable use policies, as they are leveraged by powerful institutions as a mode of control over their subjects. Enforcing acceptable use policies often requires developers to monitor users' interactions with foundation models closely, which could facilitate privacy breaches if data protection is inadequate (Wachter and Mittelstadt 2019; Kayes and Iamnitchi 2017).

Potential Negative Externalities of Enforcement

Restricting Researcher Access Longpre et al. (2024b) find that of 7 major developers with acceptable use policies, none provide comprehensive exemptions for researchers. Platforms that distribute foundation models may rate limit or ban accounts that violate acceptable use policies, even if those accounts belong to researchers, meaning that acceptable use policies can act as a disincentive against carrying out adversarial red teaming. Concerns regarding restrictions on researcher access led over 350 researchers and advocates to sign an open letter calling for companies to offer safe harbor and refrain from disproportionate enforcement of their acceptable use policies in such cases.

Case Studies of AUPs Preventing Beneficial Uses Strict acceptable use policies can inadvertently prevent a wide variety of beneficial uses of foundation models. Acceptable use policies do not permit users to generate prohibited content when doing so would likely be net beneficial in a specific context or circumstance, meaning that they function as a blanket ban on certain types of content (Small 2023).

Acceptable use policies can ban entire domains of use, but this may be overly cautious in scoping out applications. For example, Meta's acceptable use policy for Llama 2 states "You agree you will not use, or allow others to use, Llama 2 to...Engage in, promote, incite, facilitate, or assist in the planning or development of activities that present a risk of death or bodily harm to individuals, including use of Llama 2 related to...Operation of critical infrastructure, transportation technologies, or heavy machinery". Critical infrastructure and heavy machinery are not defined in the policy, making this restriction expansive. If a robotics company were to use Llama 2 to assist in turning transcribed audio instructions into commands for a robot, Meta could claim that the company had violated its prohibition against using Llama 2 to assist with heavy machinery. Language models are used in numerous ways in robotics research, and acceptable use policies could limit such work (Kim et al. 2024).

Acceptable use policies can also prevent the use of models to generate content that developers consider obscene, even when it could be beneficial (Stardust 2018; Jones 2020; Bronstein 2021). In preventing generation of sexual content, an acceptable use policy would prohibit the use of a foundation model to assist in reducing harm associated with sex work (Bernier et al. 2021; Sanders et al. 2018; Rekart

2005). Sex workers would be prevented from using chatbots to respond to their clients, though this might reduce the amount of harassment to which they are exposed (Hamilton, Barakat, and Redmiles 2022). Sex workers would also not be able to create consensual intimate images, which may distort the market for images of their likenesses such that it will be dominated by non-consensual intimate images rather than images they create themselves (Cole 2023).

In a similar vein, restrictions on generating content related to illicit substances may undermine harm reduction initiatives (Ezell et al. 2024). Character.AI's acceptable use policy states that "[y]ou agree not to submit any Content that...seeks to buy or sell illegal drugs", while four other developers' policies prohibit content related to illicit substances (Anthropic, Meta, Google, OpenAI). These restrictions impact not only organized criminal groups seeking to scale-up mass distribution of illicit substances, but also social services organizations that follow best practices by promoting harm reduction rather than abstinence for populations with substance use disorders (SAMHSA 2023).

These potentially beneficial uses of generating prohibited content should lead developers to weigh the costs and benefits of including and enforcing each prohibited use in their policies. Some developers may choose to not enforce their policies in risky domains that could present benefits (e.g., robotics and harm reduction), making their policies less stringent in practice. But many developers reuse policies from other organizations, promoting standardization while reducing the likelihood that each provision will be carefully considered. Marginalized populations such as sex workers may be harmed by disproportionate policy enforcement.

Lack of Transparency in Enforcement

There is little publicly available information about how acceptable use policies are enforced (Shahi, Conner, and Alvarez 2024). Although firms publish the prohibited uses of their models, it is unclear how they enforce their policies in practice. Foundation model developers provide little or no information about how they respond to policy violations, or whether they provide justification or appeals processes when they do so (Bommasani et al. 2023b). Bommasani and Klyman et al. (2024) compile transparency reports from foundation model developers, finding that 10 of 14 disclosed some high-level details related to enforcement, though just 8 disclosed if they allow users to appeal decisions and 7 disclosed if justification is provided when enforcement occurs; notably, Google disclosed it has not taken any enforcement actions under its Generative AI Prohibited Use Policy.

This lack of transparency is different from other digital technologies; social media companies, for instance, regularly release transparency reports providing details about enforcement of acceptable use policies and other provisions in their TOS (Kaushal et al. 2024). Still, as Douek (2022) writes, "it's hard to overstate both how ineffective platforms are at enforcing their rules, and how little is known about what systems they have in place to do so." Companies are moving quickly to deploy foundation models at the same time as they downsize the trust and safety teams required to enforce acceptable use policies (Motyl and Ellingson 2024).

Without information about how acceptable use policies are enforced, it is not evident that they are currently being implemented or effective in limiting dangerous uses (Bommasani et al. 2024). Some firms may publish acceptable use policies as a type of public relations statement to show they are responsible organizations, as they incur no costs for doing so if they do not invest in enforcement (Floridi 2019).

Discussion

Developers Decide What Uses Are Acceptable

Acceptable use policies are written by developers without input from users or external partners. Developers alone have the ability to decide how their foundation models and the AI systems that integrate them are used; foundation models sit at the center of generative AI supply chain, granting developers outsized power in this ecosystem (Bommasani et al. 2023c). While corporate users may negotiate more permissive terms, individuals have no means of negotiating changes to the terms of service. Foundation model developers with acceptable use policies include some of the world's largest companies (e.g., Amazon, Google, Meta), and their choice of what constitutes unacceptable use stems in large part from the need to reduce the legal, political, and reputational risks they experience, not risks to their users (Gillespie 2018).

Since 2023, developers have made some effort to broaden the group of people responsible for determining the boundaries of acceptable use. For instance, Huang et al. (2024) conducted a survey of Americans to solicit their views regarding how language models should behave, then updated the model behavior policy for an Anthropic model by using respondents' preference data during fine-tuning. This is part of what Delgado et al. (2023) call the "participatory turn in AI design," with some developers suggesting they may incorporate surveys into policy development (Suresh et al. 2024; Birhane et al. 2022). Open-weight foundation models without use restrictions also widen the circle of who can be involved in such decisions, allowing downstream developers to choose acceptable use policies (Bommasani et al. 2023a).

But these efforts to broaden participation in policy design fall short of addressing the lack of legitimacy that firms may face in deciding how an entire class of new general-purpose technologies may be used (Suresh et al. 2024; Cooper and Zafiroglu 2024; Widder 2024). Technology companies were not chosen to be the arbiters of what AI-generated content is acceptable by a democratic process (Gautam 2024; Seger et al. 2023); rather, as Ovadya (2023) writes, powerful corporations that "unilaterally control extraordinarily powerful AI systems" may represent a form of "autocratic centralization." Several of the largest foundation model developers are currently facing lawsuits which allege they broke antitrust law to obtain their dominant market position (D.D.C. 2024). The oligopoly in the cloud market limits the ability of startups and competitors to develop and distribute foundation models without the influence of incumbents, further concentrating decision-making power over what constitutes acceptable use (CMA 2024; Hu, Bensinger, and Godoy 2024).

Developers' enforcement of acceptable use policies for foundation models is likely to suffer from many of the is-

sues digital platforms face in enforcing their content policies (Gillespie 2018). Social media companies are regularly accused of disparate and unequal enforcement of their policies, amplifying white supremacist, misogynist, and far-right content while enforcing their policies against Muslims, people of color, and dissidents (Haimson et al. 2021; Siapera and Viejo-Otero 2021; Donovan, Dreyfuss, and Friedberg 2022). Marginalized communities have fewer resources for advocacy to persuade firms that their content should be considered "acceptable," meaning that centralized decision-making regarding policy enforcement often reinforces majoritarian views (Solaiman et al. 2024).

Gaps in Use Restrictions May Facilitate Misuse

Developers' acceptable use policies have substantial differences in key areas. While many developers restrict content related to politics and medical advice, more than two-thirds of developers have no such prohibitions. And while some companies' policies prevent their models from being used by content farms or the legal services industry, some have few industry-related restrictions and others release noncommercial models with no other restricted categories of use.

The lack of consistency across developers' acceptable use policies could facilitate misuse in three ways. First, it makes policy enforcement more difficult. Different policies may require different enforcement mechanisms; for example, building a filter for prompts related to glorifying violence requires different data (e.g., word blocklists) than for prompts related to producing malware (Jhaver et al. 2018). As a result, it is more difficult for deployers to enforce the acceptable use policies for models on their platforms, creating opportunities for deliberate misuse. It is also unclear how to properly combine two acceptable use policies for different models (Villa 2022), as would be needed in the case of a model that makes use of other models, as with model merging (Choshen et al. 2022), mixture-of-agents (Wang et al. 2024b), or other systems in which an agent interacts with other models (Lai et al. 2024). And if the outputs of a model are used as part of the training data of another model, the latter model might include data that does not reflect its acceptable use policy if the two models' AUPs differ.

Second, the lack of consistency diminishes users' understanding of what uses of a foundation model are acceptable. Many users regularly interact with multiple foundation models, such as the voice assistant on a smartphone, the summarization model in a search engine, and a standalone chatbot for brainstorming or coding. Each of these models may have a different acceptable use policy, meaning the average user may struggle to internalize which uses are disallowed.

Third, models are not safety-tuned for less common restricted uses. Without strong norms in the developer community about which uses are unacceptable, developers are less likely to invest in making their models refuse to generate related content (Reuel et al. 2024). As a result, there is a lack of data that developers can use to build filters for less common prohibited use categories, such as self-harm.

Every acceptable use policy need not be the same, but the lack of standardization is creating negative externalities in the ecosystem. At minimum, developers could work to build consensus around what constitutes acceptable use and aim to make their policies interoperable where appropriate.

AUPs Shape the Foundation Model Market

AUPs alter the foundation model market by affecting which organizations can use a model and for what purpose. For example, developers use these policies to prevent companies from making use of their services, stealing their intellectual property, or building a competing model. 24 companies ban firms and other users from using their models to train other machine learning models, restricting the supply of datasets of model outputs and concentrating the market for models that are trained on their models' outputs (Zhao et al. 2024). On the other hand, in July 2024 Meta updated a license to allow users to use outputs from Llama 3.1 to "to create, train, fine tune, or otherwise improve an AI model," perhaps in an effort to gain market share (Lambert 2024).

Acceptable use policies also help determine what industries can make use of developers' models. Policies that prohibit the use of models for weapons production may block the arms industry from making use of those foundation models, as with surveillance tech companies and political advocacy groups. These policies determine the types of uses of models that are permitted as well (e.g., no automated decision systems, no automated posting of AI-generated outputs). Even industries that are allowed to make use of models may not be able to do so for common applications.

Areas for Future Work

The way in which developers and deployers enforce acceptable use policies for foundation models remains opaque. Collecting data related to enforcement is a key area for future work, as there is little indication that companies will share enforcement data (Bommasani et al. 2024). This data might be collected by asking users to donate their data (e.g., chat logs), surveying users about their experiences, or working with companies to gain access. One key question is how enforcement differs based on the system the foundation model is embedded within; for instance, some companies might enforce their acceptable use policy less strictly for language models distributed via API as opposed to via a chat interface, as there are more users of chatbots.

Content moderation has been studied more thoroughly on social media platforms than on generative AI platforms despite researchers' access to foundation models and lack of access to underlying recommender systems (Mahomed et al. 2024). Some generative AI startups have adopted content moderation practices quickly, hiring trust and safety teams (often from incumbents) and adopting acceptable use policies to curb undesirable content. The same scrutiny that is applied to content moderation on social media should be applied to developers' enforcement of acceptable use policies, including the data labor employed as part of this work (Gray and Suri 2019; Roberts 2019). Evaluations of foundation models can also be seen as a form of content moderation, as they are used to assess whether a model will produce violative content and inform interventions to reduce this behavior (Nangia et al. 2020; Zhao et al. 2018).

Acknowledgments

I thank Ahmed Ahmed, Sanna Ali, Rishi Bommasani, Peter Cihon, Evelyn Douek, Carlos Muñoz Ferrandis, Peter Henderson, Daniel Ho, Aspen Hopkins, Sayash Kapoor, Percy Liang, Shayne Longpre, Emma Lurie, Daniel McDuff, Aviya Skowron, Dave Willner, Betty Xiong, and Yi Zeng for feedback and discussions on this work. All errors and omissions are my own.

References

Ahn, J.; Bivona, L. K.; and DiScala, J. 2011. Social media access in K-12 schools: Intractable policy controversies in an evolving world. *Proceedings of the American Society for Information Science and Technology*, 48(1): 1–10.

Alfawzan, N.; Christen, M.; Spitale, G.; and Biller-Andorno, N. 2022. Privacy, Data Sharing, and Data Security Policies of Women's mHealth Apps: Scoping Review and Content Analysis. *JMIR Mhealth Uhealth*, 10(5): e33735.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Barnett, J. 2023. The Ethical Implications of Generative Audio Models: A Systematic Literature Review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23. ACM.

Basdevant, A.; François, C.; Storchan, V.; Bankston, K.; Bdeir, A.; Behlendorf, B.; Debbah, M.; Kapoor, S.; LeCun, Y.; Surman, M.; King-Turvey, H.; Lambert, N.; Maffulli, S.; Marda, N.; Shivkumar, G.; and Tunney, J. 2024. Towards a Framework for Openness in Foundation Models: Proceedings from the Columbia Convening on Openness in Artificial Intelligence. arXiv:2405.15802.

Bernier, T.; Shah, A.; Ross, L. E.; Logie, C. H.; and Seto, E. 2021. The Use of Information and Communication Technologies by Sex Workers to Manage Occupational Health and Safety: Scoping Review. *Journal of Medical Internet Research*, 23(6): e26085.

Beyer, K. E. 2014. Busting the Ghost Guns: A Technical, Statutory, and Practical Approach to the 3-D Printed Weapon Problem. *Kentucky Law Journal*, 103: 433–456.

Biddle, S. 2024. OpenAI Quietly Deletes Ban on Using ChatGPT for "Military and Warfare". *The Intercept*.

Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394772.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.; Chen, A.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X. L.; Li, X.; Ma, T.; Malik, A.; Manning, C. D.; Mirchandani, S.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie, A.; Niebles, J. C.; Nilforoshan, H.; Nyarko, J.; Ogut, G.; Orr, L.; Papadimitriou, I.; Park, J. S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y.; Ruiz, C.; Ryan, J.; Ré, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K.; Tamkin, A.; Taori, R.; Thomas, A. W.; Tramèr, F.; Wang, R. E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S. M.; Yasunaga, M.; You, J.; Zaharia, M.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; and Liang, P. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258.

Bommasani, R.; Kapoor, S.; Klyman, K.; Longpre, S.; Ramaswami, A.; Zhang, D.; Schaake, M.; Ho, D. E.; Narayanan, A.; and Liang, P. 2023a. Considerations for Governing Open Foundation Models. Technical report, Stanford Institute for Human-Centered AI.

Bommasani, R.; Klyman, K.; Longpre, S.; Kapoor, S.; Maslej, N.; Xiong, B.; Zhang, D.; and Liang, P. 2023b. The Foundation Model Transparency Index. arXiv:2310.12941.

Bommasani, R.; Klyman, K.; Longpre, S.; Xiong, B.; Kapoor, S.; Maslej, N.; Narayanan, A.; and Liang, P. 2024. Foundation Model Transparency Reports. arXiv:2402.16268.

Bommasani, R.; Soylu, D.; Liao, T. I.; Creel, K. A.; and Liang, P. 2023c. Ecosystem Graphs: The Social Footprint of Foundation Models. arXiv:2303.15772.

Bommasani and Klyman; Kapoor, S.; Longpre, S.; Xiong, B.; Maslej, N.; and Liang, P. 2024. The Foundation Model Transparency Index v1.1: May 2024. arXiv:2407.12929.

Brenes, M.; and Hartung, W. D. 2024. Private Finance and the Quest to Remake Modern Warfare. Research report, Quincy Institute for Responsible Statecraft.

Bronstein, C. 2021. Deplatforming sexual speech in the age of FOSTA/SESTA. *Porn Studies*, 8(4): 367–380.

CAC. 2023. Interim Measures for the Management of Generative Artificial Intelligence Services. https://www.chinalawtranslate.com/en/generative-ai-interim/. Accessed: 2024-05-14.

Cen, S. H.; Hopkins, A.; Ilyas, A.; Madry, A.; Struckman, I.; and Videgaray Caso, L. 2023. AI Supply Chains.

Chandra, B.; Awad, G.; Lee, Y.; Fontana, P.; Amironesei, R.; Przybocki, M.; Roberts, K.; Tabassi, E.; Heyman, M.; and Dunietz, J. 2024. Reducing Risks Posed by Synthetic Content.

Chi, N.; Lurie, E.; and Mulligan, D. K. 2021. Reconfiguring Diversity and Inclusion for AI Ethics. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 447–457. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.

Choshen, L.; Venezian, E.; Slonim, N.; and Katz, Y. 2022. Fusing finetuned models for better pretraining. arXiv:2204.03044.

CIP; Digital Forensic Research Lab; Graphika; and Stanford Internet Observatory. 2021. The Long Fuse: Misinformation and the 2020 Election. Stanford Digital Repository: Election Integrity Partnership. v1.3.0.

CMA. 2024. AI Foundation Models: Technical update report. Technical report, UK Competition Markets Authority.

Cohere; OpenAI; and AI21 Labs. 2022. Best practices for deploying Language Models.

Cole, S. 2023. Riley Reid on AI: 'I Don't Want Porn to Get Left Behind'.

Contractor, D.; McDuff, D.; Haines, J. K.; Lee, J.; Hines, C.; Hecht, B.; Vincent, N.; and Li, H. 2022. Behavioral Use Licensing for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. ACM.

Cooper, A. F.; Moss, E.; Laufer, B.; and Nissenbaum, H. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 864–876. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Cooper, N.; and Zafiroglu, A. 2024. From Fitting Participation to Forging Relationships: The Art of Participatory ML. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

D.D.C. 2024. United States et al. v. Google LLC. Case No. 20-cv-3010 (APM), Memorandum Opinion.

Delgado, F.; Yang, S.; Madaio, M.; and Yang, Q. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703812.

Doherty, N.; Anastasakis, L.; and Fulford, H. 2011. Reinforcing the security of corporate information resources: A critical review of the role of the acceptable use policy. *International Journal of Information Management*, 31: 201–209.

Donovan, J.; Dreyfuss, E.; and Friedberg, B. 2022. *Meme Wars: The Untold Story of the Online Battles Upending Democracy in America*. Bloomsbury Publishing. ISBN 9781635578645.

Douek, E. 2022. Content Moderation as Systems Thinking. *Harvard Law Review*.

Downing, K. 2023. AI Licensing Can't Balance "Open" with "Responsible".

Downing, K. 2024. Choose Your Own Adventure: The EU AI Act and Openish AI.

Elo, S.; and Kyngäs, H. 2008. The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1): 107–115.

EU. 2024. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Accessed: 2024-05-14.

Ezell, J. M.; Ajayi, B. P.; Parikh, T.; Miller, K.; Rains, A.; and Scales, D. 2024. Drug Use and Artificial Intelligence: Weighing Concerns and Possibilities for Prevention. *American Journal of Preventive Medicine*, 66(3): 568–572.

Feldstein, S. 2019. *The global expansion of AI surveillance*, volume 17. Carnegie Endowment for International Peace Washington, DC.

Fergusson, G.; Fitzgerald, C.; Frascella, C.; Iorio, M.; McBrien, T.; Schroeder, C.; Winters, B.; and Zhou, E. 2023. Generating Harms: Generative AI's Impact & Paths Forward. Technical report, Electronic Privacy Information Center.

Ferretti, T. 2022. An Institutionalist Approach to AI Ethics: Justifying the Priority of Government Regulation over Self-Regulation. *Moral Philosophy and Politics*, 9(2): 239–265.

Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A.; and Srikumar, M. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *Berkman Klein Center Research Publication*, (2020-1). Http://dx.doi.org/10.2139/ssrn.3518482.

Floridi, L. 2019. Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology*, 32: 185–193.

Gautam, A. 2024. Reconfiguring Participatory Design to Resist AI Realism. *arXiv preprint arXiv:2406.03245*. Presented at Participatory Design Conference 2024.

Gillespie, T. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press. ISBN 978-0-300-17313-0.

GitHub. 2024. GitHub Acceptable Use Policies. Accessed: July 2024.

Goldstein, J. A.; Chao, J.; Grossman, S.; Stamos, A.; and Tomz, M. 2024. How persuasive is AI-generated propaganda? *PNAS Nexus*, 3(2): pgae034.

Goldstein, J. A.; Sastry, G.; Musser, M.; DiResta, R.; Gentzel, M.; and Sedova, K. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv:2301.04246.

Google. 2024. Policy guidelines for the Gemini app.

Gorwa, R.; and Veale, M. 2024. Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries. arXiv:2311.12573.

Gray, M.; and Suri, S. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt. ISBN 9781328566249.

Gregorio, N.; Mathanamohan, J.; Mahmoud, Q. H.; and Al-Taei, M. 2019. Hacking in the cloud. *Internet Technology Letters*, 2(1): e84.

Hacker, P. 2023. AI Regulation in Europe: From the AI Act to Future Regulatory Challenges. arXiv:2310.04072.

Hacker, P.; Cordes, J.; and Rochon, J. 2024. Regulating Gatekeeper Artificial Intelligence and Data: Transparency, Access and Fairness under the Digital Markets Act, the General Data Protection Regulation and Beyond. *European Journal of Risk Regulation*, 15(1): 49–86.

Haimson, O. L.; Delmonaco, D.; Nie, P.; and Wegner, A. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–35.

Hamilton, V.; Barakat, H.; and Redmiles, E. M. 2022. Risk, Resilience and Reward: Impacts of Shifting to Digital Sex Work. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Henderson, P.; Li, X.; Jurafsky, D.; Hashimoto, T.; Lemley, M. A.; and Liang, P. 2023. Foundation Models and Fair Use. arXiv:2303.15715.

Henderson, P.; Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; and Mittal, P. 2024. Safety Risks from Customizing Foundation Models via Fine-tuning. Policy brief, Stanford Institute for Human-Centered AI.

Hoffmann, M.; and Frase, H. 2023. Adding Structure to AI Harm: An Introduction to CSET's AI Harm Framework. Technical report, Center for Security and Emerging Technology.

Hu, K.; Bensinger, G.; and Godoy, J. 2024. Exclusive: FTC seeking details on Amazon deal with AI startup Adept, source says. *Reuters*. Accessed [Insert Access Date].

Huang, S.; Siddarth, D.; Lovitt, L.; Liao, T. I.; Durmus, E.; Tamkin, A.; and Ganguli, D. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1395–1417. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Hugging Face. 2023. Content Policy.

Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; and Khabsa, M. 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. arXiv:2312.06674.

Jhaver, S.; Ghoshal, S.; Bruckman, A.; and Gilbert, E. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2): 1–33.

Jones, A. 2020. *Camming: Money, Power, and Pleasure in the Sex Work Industry*. NYU Press. ISBN 9781479815470.

Jones, B. M. 2015. 3D Printing in Libraries: A View from Within the American Library Association: Privacy, Intellectual Freedom and Ethical Policy Framework. *Bulletin of the Association for Information Science and Technology*, 42(1): 36–41.

Kapoor, S.; Bommasani, R.; Klyman, K.; Longpre, S.; Ramaswami, A.; Cihon, P.; Hopkins, A.; Bankston, K.; Biderman, S.; Bogen, M.; Chowdhury, R.; Engler, A.; Henderson, P.; Jernite, Y.; Lazar, S.; Maffulli, S.; Nelson, A.; Pineau, J.; Skowron, A.; Song, D.; Storchan, V.; Zhang, D.; Ho, D. E.; Liang, P.; and Narayanan, A. 2024. On the Societal Impact of Open Foundation Models. arXiv:2403.07918.

Kaushal, R.; van de Kerkhof, J.; Goanta, C.; Spanakis, G.; and Iamnitchi, A. 2024. Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database. arXiv:2404.02894.

Kayes, I.; and Iamnitchi, A. 2017. Privacy and security in online social networks: A survey. *Online Social Networks and Media*, 3: 1–21.

Keller, D. 2021. Amplification and its discontents: Why regulating the reach of online content is hard. *J. Free Speech L.*, 1: 227.

Keller, P.; and Bonato, N. 2023. Growth of responsible AI licensing. Analysis of license use for ML models published on. *Open Future*.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; Vuong, Q.; Kollar, T.; Burchfiel, B.; Tedrake, R.; Sadigh, D.; Levine, S.; Liang, P.; and Finn, C. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. arXiv:2406.09246.

Lai, S.; Potter, Y.; Kim, J.; Zhuang, R.; Song, D.; and Evans, J. 2024. Position: Evolving AI Collectives Enhance Human Diversity and Enable Self-Regulation. In *Forty-first International Conference on Machine Learning*.

Lambert, N. 2024. Llama 3.1 405B, Meta's AI strategy, and the new, open frontier model ecosystem. *Interconnects*.

Li, H.; Zhang, J.; and Sarathy, R. 2010. Understanding compliance with internet use policy from the perspective of rational choice theory. *Decision Support Systems*, 48(4): 635–645.

Longpre, S.; Biderman, S.; Albalak, A.; Schoelkopf, H.; McDuff, D.; Kapoor, S.; Klyman, K.; Lo, K.; Ilharco, G.; San, N.; Rauh, M.; Skowron, A.; Vidgen, B.; Weidinger, L.; Narayanan, A.; Sanh, V.; Adelani, D.; Liang, P.; Bommasani, R.; Henderson, P.; Luccioni, S.; Jernite, Y.; and Soldaini, L. 2024a. The Responsible Foundation Model Development Cheatsheet: A Review of Tools Resources. arXiv:2406.16746.

Longpre, S.; Kapoor, S.; Klyman, K.; Ramaswami, A.; Bommasani, R.; Blili-Hamelin, B.; Huang, Y.; Skowron, A.; Yong, Z.-X.; Kotha, S.; Zeng, Y.; Shi, W.; Yang, X.; Southen, R.; Robey, A.; Chao, P.; Yang, D.; Jia, R.; Kang, D.; Pentland, S.; Narayanan, A.; Liang, P.; and Henderson,

P. 2024b. A Safe Harbor for AI Evaluation and Red Teaming. arXiv:2403.04893.

Looney, S. 2023. Content moderation through removal of service: Content delivery networks and extremist websites. *Policy & Internet*, 15(4): 544–558.

Mahomed, Y.; Crawford, C. M.; Gautam, S.; Friedler, S. A.; and Metaxa, D. 2024. Auditing GPT's Content Moderation Guardrails: Can ChatGPT Write Your Favorite TV Show? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 660–686. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Marchal, N.; Xu, R.; Elasmar, R.; Gabriel, I.; Goldberg, B.; and Isaac, W. 2024. Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data. arXiv:2406.13843.

Maus, N.; Chao, P.; Wong, E.; and Gardner, J. R. 2023. Black box adversarial prompting for foundation models. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*.

Mayring, P. 2015. *Qualitative Content Analysis: Theoretical Background and Procedures*, 365–380. Dordrecht: Springer Netherlands. ISBN 978-94-017-9181-6.

McDuff, D.; Korjakow, T.; Cambo, S.; Benjamin, J. J.; Lee, J.; Jernite, Y.; Ferrandis, C. M.; Gokaslan, A.; Tarkowski, A.; Lindley, J.; Cooper, A. F.; and Contractor, D. 2024. On the Standardization of Behavioral Use Clauses and Their Adoption for Responsible Licensing of AI. arXiv:2402.05979.

McMenemy, D. 2019. *Public Library Digital Services: Emergent Issues of Access and Acceptable Use*.

Metz, R.; and Ford, B. 2024. Adobe's 'Ethical' Firefly AI Was Trained on Midjourney Images. *Bloomberg*.

Minow, M.; Lipinski, T. A.; McCord, G.; et al. 2016. *The Library's Legal Answers for Makerspaces*. ALA Editions. ISBN 978-0-8389-1390-1. EBook.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 220–229. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Motyl, M.; and Ellingson, G. 2024. The Unbearably High Cost of Cutting Trust Safety Corners.

Mouton, C. A.; Lucas, C.; and Guest, E. 2024. *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study*. Santa Monica, CA: RAND Corporation.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. arXiv:2010.00133.

Ng, D. T. K.; Leung, J. K. L.; Chu, S. K. W.; and Qiao, M. S. 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2: 100041.

O'Byrne, W. I. 2019. *Acceptable Use Policies*, 1–6. John Wiley Sons, Ltd. ISBN 9781118978238.

OpenAI. 2024. Model Spec.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Łukasz Kaiser; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Łukasz Kondraciuk; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Malfacini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin, B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.; McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta, A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.; Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.; Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Neelakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O'Keefe, C.; Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo, G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.; Peng, A.; Perelman, A.; de Avila Belbute Peres, F.; Petrov, M.; de Oliveira Pinto, H. P.; Michael; Pokorny; Pokrass, M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl, E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond, C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez, H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.; Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Selsam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker, S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.; Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher, N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak, N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng, E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Vallone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang, J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C.; Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff, M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Work-

man, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo, S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.; Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; and Zoph, B. 2024. GPT-4 Technical Report. arXiv:2303.08774.

OSI. 2024. The Open Source AI Definition – draft v. 0.0.8. Accessed July 2024.

Ovadya, A. 2023. Reimagining Democracy for AI. *Journal of Democracy*, 34(4): 162–170.

Pater, J. A.; Kim, M. K.; Mynatt, E. D.; and Fiesler, C. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 2016 ACM International Conference on Supporting Group Work*, GROUP '16, 369–374. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342766.

Pfefferkorn, R. 2024. Addressing Computer-Generated Child Sex Abuse Imagery: Legal Framework and Policy Implications. *Lawfare*. Accessed [Insert Access Date].

Pistilli, G.; Muñoz Ferrandis, C.; Jernite, Y.; and Mitchell, M. 2023. Stronger Together: on the Articulation of Ethical Charters, Legal Tools, and Technical Documentation in ML. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 343–354. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! arXiv:2310.03693.

Rekart, M. L. 2005. Sex-work harm reduction. *The Lancet*, 366(9503): 2123–2134.

Reuel, A.; Bucknall, B.; Casper, S.; Fist, T.; Soder, L.; Aarne, O.; Hammond, L.; Ibrahim, L.; Chan, A.; Wills, P.; Anderljung, M.; Garfinkel, B.; Heim, L.; Trask, A.; Mukobi, G.; Schaeffer, R.; Baker, M.; Hooker, S.; Solaiman, I.; Luccioni, A. S.; Rajkumar, N.; Moës, N.; Ladish, J.; Guha, N.; Newman, J.; Bengio, Y.; South, T.; Pentland, A.; Koyejo, S.; Kochenderfer, M. J.; and Trager, R. 2024. Open Problems in Technical AI Governance. arXiv:2407.14981.

Roberts, S. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press. ISBN 978-0-300-24531-8.

Robinson, E. 2019. *The panoptic principle: privacy and surveillance in the public library as evidenced in the acceptable use policy*. Thesis, University of Strathclyde. Accessed: 2021-07-01.

Robinson, E.; and McMenemy, D. 2020. 'To be understood as to understand': A readability analysis of public library acceptable use policies. *Journal of Librarianship and Information Science*, 52(3): 713–725.

SAMHSA. 2023. Harm Reduction Framework. Technical report, Center for Substance Abuse Prevention, Substance Abuse and Mental Health Services Administration.

Sanders, T.; Scoular, J.; Campbell, R.; Pitcher, J.; and Cunningham, S. 2018. *Internet sex work: Beyond the gaze*. Springer.

Seger, E.; Ovadya, A.; Garfinkel, B.; Siddarth, D.; and Dafoe, A. 2023. Democratising AI: Multiple Meanings, Goals, and Methods. arXiv:2303.12642.

Shah, R.; Montixi, Q. F.; Pour, S.; Tagade, A.; and Rando, J. 2023. Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation. In *Socially Responsible Language Modelling Research*.

Shahi, M.; Conner, A.; and Alvarez, N. 2024. Generative AI Should Be Developed and Deployed Responsibly at Every Level for Everyone.

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla, N.; Gallegos, J.; Smart, A.; Garcia, E.; and Virk, G. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. arXiv:2210.05791.

Siapera, E.; and Viejo-Otero, P. 2021. Governing hate: Facebook and digital racism. *Television & New Media*, 22(2): 112–130.

Siau, K.; Nah, F. F.-H.; and Teng, L. 2002. Acceptable internet use policy. *Commun. ACM*, 45(1): 75–79.

Small, Z. 2023. Black Artists Say A.I. Shows Bias, With Algorithms Erasing Their History. *The New York Times*.

Solaiman, I. 2023. The Gradient of Generative AI Release: Methods and Considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 111–122. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Solaiman, I.; Talat, Z.; Agnew, W.; Ahmad, L.; Baker, D.; Blodgett, S. L.; Chen, C.; au2, H. D. I.; Dodge, J.; Duan, I.; Evans, E.; Friedrich, F.; Ghosh, A.; Gohar, U.; Hooker, S.; Jernite, Y.; Kalluri, R.; Lusoli, A.; Leidinger, A.; Lin, M.; Lin, X.; Luccioni, S.; Mickel, J.; Mitchell, M.; Newman, J.; Ovalle, A.; Png, M.-T.; Singh, S.; Strait, A.; Struppek, L.; and Subramonian, A. 2024. Evaluating the Social Impact of Generative AI Systems in Systems and Society. arXiv:2306.05949.

Srikumar, M.; Chang, J.; Chmielinski, K.; and on AI, P. 2024. Risk Mitigation Strategies for the Open Foundation Model Value Chain.

Stardust, Z. 2018. Safe for Work: Feminist Porn, Corporate Regulation and Community Standards. In Dale, C.; and Overell, R., eds., *Orienting Feminism: Media, Activism and Cultural Representation*, 155–179. Cham: Springer International Publishing.

Stewart, F. 2000. Internet Acceptable Use Policies: Navigating the Management, Legal, and Technical Issues. *Information Systems Security*, 9(3): 1–7.

Suchman, L. 2020. Algorithmic warfare and the reinvention of accuracy. *Critical Studies on Security*, 8(2): 175–187.

Suresh, H.; Tseng, E.; Young, M.; Gray, M.; Pierson, E.; and Levy, K. 2024. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1609–1621. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Thorn. 2024. Safety by Design for Generative AI: Preventing Child Sexual Abuse.

Villa, L. 2022. Evaluating the RAIL license family.

Wachter, S.; and Mittelstadt, B. 2019. A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Colum. Bus. L. Rev.*, 494.

Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; Truong, S. T.; Arora, S.; Mazeika, M.; Hendrycks, D.; Lin, Z.; Cheng, Y.; Koyejo, S.; Song, D.; and Li, B. 2024a. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. arXiv:2306.11698.

Wang, J.; Wang, J.; Athiwaratkun, B.; Zhang, C.; and Zou, J. 2024b. Mixture-of-Agents Enhances Large Language Model Capabilities. arXiv:2406.04692.

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2023. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; Gabriel, I.; Rieser, V.; and Isaac, W. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986.

Weidman, J.; and Grossklags, J. 2019. The Acceptable State: An Analysis of the Current State of Acceptable Use Policies in Academic Institutions. In *Proceedings of the 27th European Conference on Information Systems (ECIS)*, Research Papers. Stockholm & Uppsala, Sweden. ISBN 978-1-7336325-0-8.

White, M.; Haddad, I.; Osborne, C.; Yanglet, X.-Y. L.; Abdelmonsef, A.; and Varghese, S. 2024. The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence. arXiv:2403.13784.

White House. 2023. Voluntary AI Commitments. Accessed: 2024-05-14.

Widder, D. G. 2024. Epistemic Power in AI Ethics Labor: Legitimizing Located Complaints. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1295–1304. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Yang, X.; Wang, X.; Zhang, Q.; Petzold, L.; Wang, W. Y.; Zhao, X.; and Lin, D. 2023. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*.

Zeng, Y.; Yang, Y.; Zhou, A.; Tan, J. Z.; Tu, Y.; Mai, Y.; Klyman, K.; Pan, M.; Jia, R.; Song, D.; Liang, P.; and Li, B. 2024. AIR-Bench 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies. arXiv:2407.17436.

Zeng and Klyman; Zhou, A.; Yang, Y.; Pan, M.; Jia, R.; Song, D.; Liang, P.; and Li, B. 2024. AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies. arXiv:2406.17864.

Zhan, Q.; Fang, R.; Bindu, R.; Gupta, A.; Hashimoto, T.; and Kang, D. 2023. Removing RLHF Protections in GPT-4 via Fine-Tuning. *arXiv preprint arXiv:2311.05553*.

Zhang, A. H. 2024. The Promise and Perils of China's Regulation of Artificial Intelligence. *University of Hong Kong Faculty of Law Research Paper*, 2024(02). Http://dx.doi.org/10.2139/ssrn.4708676.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. arXiv:1804.06876.

Zhao, W.; Ren, X.; Hessel, J.; Cardie, C.; Choi, Y.; and Deng, Y. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. arXiv:2405.01470.

Zou, A.; Wang, Z.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.