Acceptable Use Policies for Foundation Models

Kevin Klyman
Stanford University, Center for Research on Foundation Models Harvard University, Belfer Center for Science and International Affairs 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. Developers 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. Nevertheless, 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 overall AI ecosystem.

1. Introduction

Policymakers hoping to regulate foundation models have focused on preventing specific objectionable uses of AI systems, such as the creation of bioweapons [113], deepfakes [25], and child sexual abuse material [159]. 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 [12]. Yet developers of foundation models have been proactive, adopting broad policies as part of their terms of service or model licenses that prohibit many potentially dangerous uses of the technology.

Foundation model developers have taken several approaches to adopting legally binding use restrictions. [106] find that developers of open-weight foundation models increasingly distribute these models with licenses that include a standardized set of behavioral use restrictions. Developers of closed-weight models have also restricted how users can make use of their models, often via terms of service agreements that prohibit generating specific categories of content [17]. 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: Section 2 provides background on acceptable use policies for foundation models, comparing them to similar policies for other technologies and to documents like model cards (which list out-of-scope uses but are not legally binding). Section 3 describes the methodology used to identify acceptable use policies and analyze their content. Section 4 analyzes the differences between developers' policies in terms of prohibited content and restrictions on types of end use. Section 5 outlines difficulties in policy enforcement and potential downsides from strict enforcement. Section 6 discusses developers' decision-making power, how gaps in use restrictions may facilitate misuse, and how acceptable use policies shape the foundation model market. Section 7 identifies areas for future work.

2. Background

2.1 What is an acceptable use policy?

Acceptable use policies are common across digital technologies [117]. Providers of public access computers [133], websites [151, 172], and digital platforms [123, 144] 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 [39, 134].

The acceptable use policies of social media companies [24], cloud service providers [64], and content delivery networks [97] have received scrutiny as they constrain the be-

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

First published: arXiv:2409.09041v1

havior of hundreds of millions of users. Acceptable use policies adopted by employers, which limit employees' use of company-provided technologies [93], schools, which limit students' use of the internet [1], and public libraries, which limit the public's use of public access computers [107], 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 frequently 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" [121].

In addition, many developers state that the purpose of reinforcement learning from human feedback (RLHF) is to make their foundation models less likely to generate outputs that would violate their acceptable use policies [89]. Meta's technical report for Llama 2 notes that the risks RLHF was intended to mitigate include "illicit and criminal activities (e.g., terrorism, theft, human trafficking); hateful and harmful activities (e.g., defamation, self-harm, eating disorders, discrimination); and unqualified advice (e.g., medical advice, financial advice, legal advice)," which correspond to the acceptable use policy in Llama 2's license [160]. Anthropic's model card for Claude 3 similarly says "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" [3].

2.2 How do AUPs differ from other similar 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 [112]. 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 [94].
- *Model Behavior Policies*: Model behavior policies determine what a model can or cannot do [5, 59, 120]. While acceptable use policies apply to user behavior, model behavior policies apply to the behavior of the model itself [17]. 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 [18, 28, 75]. 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 [126, 136].
- *Third party contracts*: Foundation model developers frequently partner with other firms to disseminate foundation models [23]. 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) [149]. 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.

2.3 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. [29] 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 [106].

Governments have taken an interest in acceptable use policies, which are a salient effort by foundation model developers to self-regulate [50]. 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 [65, 162]. 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 functions, and suspending or concluding the provision of services" [36, 181]. And the US

[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 [9], 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 [79, 110].

Voluntary AI Commitments require firms to publicly report "domains of appropriate and inappropriate use" as well as any limitations of the model that affect these domains [174].

Neither the EU AI Act nor the US Voluntary AI Commitments require that firms enforce their AUPs or restrict any particular uses. By contrast, China's February 2024 regulatory guidance on Basic Safety Requirements for Generative Artificial Intelligence Services specifies 31 safety risks that developers must prohibit, such as "subvert[ing] state power," "endanger[ing] national security," and "dissemination of false and harmful information" [118, 179].

3. Methodology

3.1 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 [15].
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 portion of the TOS as the AUP.
3. For each remaining developer, check the license for its "flagship foundation model";[3] if it includes behavioral use restrictions, take that portion of the license as the AUP.
4. For each remaining developer, if the TOS or license reference a separate document with behavioral use restrictions (e.g., usage guidelines) such that the restrictions are binding, take the relevant portion of that document as the AUP.

3.2 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 [105]. This was done inductively [46], with categories drawn directly from acceptable use policies, and was inspired by prior work related to AI ethics guidelines [51, 77], privacy policies [2], content moderation guidelines [24], benchmarks [167], and Responsible AI Licenses [106].

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 on GitHub), where columns show foundation model 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 [117]. 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 [4, 49, 71, 101, 141, 171], this paper assesses how companies taxonomize risk on the basis of their own policies.

4. Analysis of acceptable use policies

4.1 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, and the UAE.

4.2 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 cate-

[3] A flagship foundation model is a developer's most salient and/or capable model, informed by its public documentation [17].

[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	Model Specific Y/N (Model)	Policy Document	Flagship Model Series (Output Modality)	HQ	Openness	Ref
01.ai	Yi Series Models Community License Agreement v2.1, §2 License and License Restrictions	Y (Yi)	License	Yi (Text)	PRC	Open	[1]
Adept	Terms of Use, §1.1(d) Usage Restrictions	N	TOS	Fuyu (Multimodal)	USA	Open	[2]
Adobe	Generative AI User Guidelines	N	Standalone	Firefly (Image)	USA	Closed	[3]
AI21	Usage Guidelines	N	Standalone	Jurassic-2 (Text)	ISR	Closed	[4]
AI2	AI2 ImpACT License for Low-Risk Artifacts	Y (Tulu v2)	License	OLMo (Text)	USA	Open	[5]
Aleph Alpha	Terms and Conditions, §4.8 Customer's Rights and Use Restrictions	N	TOS	Luminous (Text)	DEU	Closed***	[6]
Amazon	AWS Responsible AI Policy & AWS Acceptable Use Policy*	N	Standalone	Titan Text (Text)	USA	Closed	[7]
Anthropic	Acceptable Use Policy	N	Standalone	Claude 3 (Text)	USA	Closed	[8]
Baidu	ERNIE Bot User Agreement, §4 Service Usage Specifications	Y (ERNIE)	TOS	ERNIE 4.0 (Text)	PRC	Closed	[9]
BigCode	BigCode Open RAIL-M v1 License, §A Use Restrictions	Y (StarCoder 2)	License	StarCoder 2 (Text)	N/A**	Open	[10]
BigScience	BigScience RAIL License v1.0, §A Use Restrictions	Y (BLOOM)	License	BLOOM (Text)	N/A**	Open	[11]
Character.AI	Terms of Service, Conditions of Use	N	TOS	Not Public (Text)	USA	Closed	[12]
Cohere	Usage Guidelines	N	Standalone	Command (Text)	CAN	Closed	[13]
Databricks	Databricks Open Model Acceptable Use Policy	Y (DBRX)	Standalone	DBRX (Text)	USA	Open	[14]
DeepSeek	Terms of Use, §3 Service Management	N	TOS	DeepSeek (Text)	PRC	Open	[15]
Eleven Labs	Terms of Service, Prohibited Activities	N	TOS	Not Public**** (Audio)	USA	Closed	[16]
Google	Generative AI Prohibited Use Policy	N	Standalone	Gemini (Multimodal)	USA	Closed	[17]
Inflection	Terms of Service, Acceptable Use	N	TOS	Inflection-2.5 (Text)	USA	Closed	[18]
Meta	Acceptable Use Policy	Y (Llama 2)	License	Llama 2 (Text)	USA	Open	[19]
Midjourney	Terms of Service, §9 Community Guidelines	N	TOS	Midjourney v6 (Image)	USA	Closed	[20]
Mistral	Terms of Use, §8 Your obligations/§9 Our Obligations & Le Chat Terms of Service, §4.3 Chat Moderation Policy*	Y (Mistral API)	TOS	Mixtral (Text)	FRA	Open	[21]
OpenAI	Usage Policy	N	Standalone	GPT-4 (Multimodal)	USA	Closed	[22]
Perplexity	Terms of Service, Acceptable Use	N	TOS	Not Public**** (Text)	USA	Closed	[23]
Reka	Terms of Service, §3.2 Responsible Use	N	TOS	Yasa-1 (Multimodal)	USA	Closed	[24]
Runway	Terms of Service, §5 User Conduct	N	TOS	Not Public**** (Video)	USA	Closed	[25]
Stability AI	Acceptable Use Policy	N	Standalone	Stable Diffusion 3 (Image)	GBR	Open	[26]
TII	Acceptable Use Policy	Y (Falcon 180B)	Standalone	Falcon 180B (Text)	UAE	Open	[27]
Together	Terms of Service, §2.4 Your Responsibilities	N	TOS	StripedHyena Nous (Text)	USA	Open	[28]
Twelve Labs	Terms of Service, §14 No Unlawful or Prohibited Use	N	TOS	Pegasus-1 (Video)	USA	Closed	[29]
Writer	Terms and Conditions, §4.3 Acceptable Use	N	TOS	Palmyra-1 (Text)	USA	Closed	[30]

Table 1: **Foundation Model Developers' Acceptable Use Policies.** This table includes information on the 30 acceptable use policies under consideration in this paper, including: the developer's name; the title of the developer's AUP and (if applicable) the section of that policy that includes use restrictions; whether the policy as applied by the developer is specific to a certain foundation model (and if so which foundation model); the type of policy document that contains the AUP (a model license, a terms of service agreement, or a standalone policy); the developer's flagship foundation model series (and the output modality of those models); the country in which the developer is headquartered; whether the weights of the flagship model series are open; and a reference to the AUP. *Amazon and Mistral's TOS explicitly refer to two relevant documents, so both are considered. **International research coalitions. ***Aleph Alpha provides model weights to customers on premises. ****These developers have not publicly disclosed the name of or details about their flagship foundation models. (Last updated April 18, 2024)

gories like political content (9), medical advice (8), weapons (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 contained in each developers' acceptable use policy and distinguishes between open- and closed-weight developers [81]. 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) reference 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 [33, 85]. 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 (see Table 2), introducing potential selection bias in computing the average. Third, unlike closed developers, open developers often cannot enforce their acceptable use 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 acceptable 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 oper-

ations 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 the fact that developers consider these to be risks associated with their models and wish to reduce their liability for such risks [92].

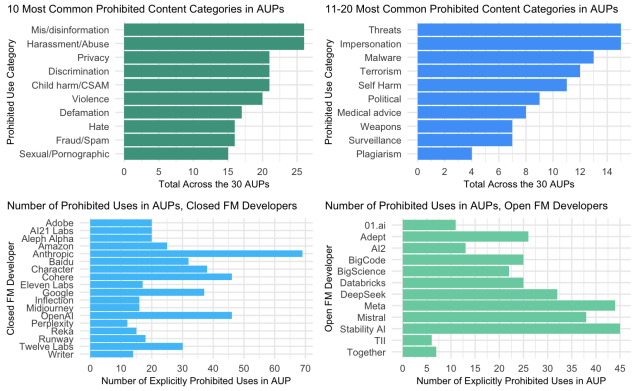

Figure 1: **Common prohibited content categories and number of prohibited uses per developer.** Top left: the 10 most common categories of content-related prohibited uses in developers' AUPs. Top right: the next 10 most common categories of content-related prohibited uses in developers' AUPs. (See the 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.

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 each of these areas, yet major developers choose not to adopt legally binding restrictions on using their models in these ways [57, 140, 154].

Other notable prohibited uses include:

- *Undermining the interests of the state*: Baidu and DeepSeek, two of three model developers in Table 1 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 Basic Safety Requirements for Generative AI Services [179].
- *Password trafficking*: Eleven Labs, the only developer in Table 1 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 [6, 101].
- *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 that some developers may be more responsive to political and reputational risk than assessments of harm or legal liability [54, 124, 158].

4.3 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 the use of their models to build a competing service, which encompasses model scraping [109].

Some developers prohibit using their models to distribute 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 [58].

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 policy 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 [19, 145]. 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" [10, 20].

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 [37, 108, 157].

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 defense and intelligence contractors respectively from making use of their foundation models [48, 80].

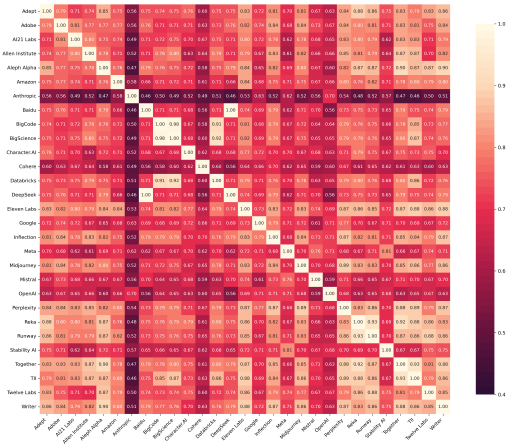

Figure 2: **Developer correlations.** The correlation between prohibited use categories for pairs of developers across all 127 categories. Correlation is measured using the simple matching coefficient (i.e. agreement rate), which is the fraction of all indicators for which both developers are assigned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

4.4 Correlations between developers' AUPs

Despite increased standardization across open developers [106], acceptable use policies 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 (the two Chinese developers) and Reka, TII, and Together (developers with relatively 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.

4.5 Developers without acceptable use policies

There are tens of developers that do not have acceptable use policies for their foundation models—table 2 provides seven examples. There are myriad reasons why a developer may choose to release a model without acceptable use policy. Some open foundation model developers do not use acceptable use policies because their models are intended for research purposes only—if they were to adopt use restrictions, it could deter researchers from conducting safety research through red teaming or adversarial attacks (in the absence of a safe harbor for good faith researchers) [7, 96]. 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 [45]. Non-commercial models such as these are frequently distributed using licenses without use restrictions such as Apache 2.0 or Creative Commons Attribution-NonCommercial licenses [95, 173]. 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 are aware of the domains in which use is permitted, while regulators and the public may be led to believe that model use is unrestricted [66, 161].

Foundation models available for commercial use may not include acceptable use policies 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 by downstream developers (e.g., Databricks' MPT-30B). Developers hoping to maximize uptake among commercial users may be less likely to adopt acceptable use policies because clients' risk-averse legal teams could 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 underdocumentation in the rush to release a model [111, 147].

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 [16].

5. Enforcement of acceptable use policies

5.1 Barriers to enforcement

5.1.1 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

Developer	Model	Intended Use (Source)	Model Assets Released	License
Alibaba Cloud	Qwen-VL	"Researchers and developers are free to use the codes and model weights of both Qwen-VL and Qwen-VL-Chat. We also allow their commercial use." (License Blurb)	Code, weights	Tongyi Qianwen License Agreement
EleutherAI	GPT-NeoX 20B	"Developed primarily for research purposes. . . . not intended for deployment as-is. It is not a product and cannot be used for human-facing interactions without supervision." (Model Card)	Data, code, weights	Apache 2.0
Meta	MusicGen-Large	"The model should not be used on downstream applications without further risk evaluation and mitigation. The model should not be used to intentionally create or disseminate music pieces that create hostile or alienating environments for people. This includes generating music that people would foreseeably find disturbing, distressing, or offensive; or content that propagates historical or current stereotypes." (Model Card)	Data, code, weights	CC-BY-NC 4.0
Microsoft	Phi-2	"Given the nature of the training data, the Phi-2 model is best suited for prompts using the QA format, the chat format, and the code format. . . . Direct adoption for production tasks without evaluation is out of scope of this project." (Model Card)	Weights	MIT
Mistral	Mixtral-8x7B	N/A (Model Card)	Code, weights	Apache 2.0
Databricks	MPT-30B	"Not intended for deployment without finetuning. It should not be used for human-facing interactions without further guardrails and user consent." (Model Card)	Code, weights	Apache 2.0
xAI	Grok-1	"Grok-1 is intended to be used as the engine behind Grok for natural language processing tasks including question answering, information retrieval, creative writing and coding assistance." (Model Card)	Weights	Apache 2.0

Table 2: **Foundation Model Developers Without Acceptable Use Policies.** 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 (and the source), the model assets released, and the license under which the model is distributed.

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 distributing the weights online such that they can be downloaded, and models are often run locally [147]. As a result, open developers have few ways of monitoring downstream use of their models, making it difficult for them to enforce their policies where models are run locally or where hosted inference is provided by another organization [149].

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 [69, 90]. [43] 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 themselves 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 [44].

On the other hand, 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 [125]. 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 unacceptable 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 [44, 119].

5.1.2 Ecosystem barriers

Another issue in gauging the enforcement of acceptable use policies is the way in which they propagate across the foundation model ecosystems. 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 have their own acceptable use policies for their platforms 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 [64]. 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 would require that for each category the deployer apply the most restrictive of its partners' acceptable use policies to every model. [60] 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 [56, 74].

These challenges are made more stark by the ease with which users can circumvent technical measures used to enforce acceptable use policies. [178] 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, [127] find that fine-tuning foundation models via an API can remove safety measures like instruction tun-

ing 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 [104, 131], jailbreaks [138, 142, 169, 184], and other methods for fine-tuning away safety measures via APIs [177, 180]. These vulnerabilities show that closed developers are likely unable to enforce their acceptable use policies in many cases [70].

5.1.3 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 [39, 172]. Acceptable use policies may be effective in limiting the behavior of corporate users, which are legally risk-averse, but are unlikely to fundamentally change the behavior of the average individual user [165].

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

Social media companies' content policies also shift responsibility for toxic content from the platform that algorithmically amplifies such content to individual users that post it [84]. 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 [26, 51]. 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 [115], the user is often not the right party to be responsible for ensuring that a foundation model is not generating violative outputs [34]. 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 foundation model for aid [62].

One solution that developers implement is increasing surveillance of their users to monitor dangerous prompts and responses. [132] 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 [83, 166].

5.2 Potential negative externalities of enforcement

5.2.1 Restricting researcher access

[96] find that of seven major foundation model 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 refrain from disproportionate enforcement of their acceptable use policies in such cases.

5.2.2 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 [146].

Acceptable use policies can ban entire domains of use, but this might 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 would plausibly have a 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 research [21, 86].

Acceptable use policies can also prevent the use of models to generate content that developers consider obscene, even when it could be beneficial [22, 78, 150]. 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 [8, 128, 135]. 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 [68]. Sex workers would also not be able to create consensual intimate images, which may inadvertently 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 [30].

In a similar vein, restrictions on generating content related to illicit substances may undermine harm reduction initiatives [47, 98]. 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 working to promote harm reduction rather than abstinence for populations with substance use disorders [102, 153].

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 acceptable use 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 acceptable use policies from other organizations, promoting standardization while reducing the likelihood that each provision will be carefully considered [106]. Marginalized populations such as sex workers may be harmed by disproportionate policy enforcement [152].

5.3 Lack of transparency in enforcement

There is little publicly available information about how acceptable use policies are enforced [139]. Although companies make the prohibited uses of their models clear, 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 [17]. [16] compile and release 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 [170].

This lack of transparency is different from other digital technologies; social media companies, for instance, regularly release transparency reports that provide details about how they enforce their acceptable use policies and other provisions in their terms of service [82]. Still, as [41] 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 have downsized the trust and safety teams required to enforce acceptable use policies [103].

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 of foundation models [14]. Some firms may publish acceptable use policies as a type of public relations statement to demonstrate they are responsible organizations, as firms incur no costs for doing so if they do not invest in enforcement [52].

6. Discussion

6.1 Developers decide what constitutes acceptable use

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 [12, 23]. While corporate users may negotiate more permissive licenses, individual users 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, Microsoft), and their choice of what constitutes unacceptable use stems in large part from the need to reduce their own legal, political, and reputational risks, not the risks to their users [55].

Since 2023, developers have made some effort to broaden the group of people responsible for determining the boundaries of acceptable use. For instance, [73] 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 [38] call the "participatory turn in AI design," with some developers suggesting they may incorporate surveys into policy development [11, 156]. Open-weight foundation models without use restrictions also widen the circle of who can be involved in such decisions, providing an opportunity for downstream developers to choose different acceptable use policies and adapt the model such that it is more likely to comply with the policy [13].

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 [35, 42, 61, 99, 156, 175]. Technology companies were not chosen to be the arbiters of what AI-generated content is acceptable by a democratic process [53, 137]; rather, as [122] 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 antitrust lawsuits in the US which allege they broke the law to obtain their dominant market position [163, 164]. 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 [31, 32, 72, 91, 176].

Developers' enforcement of acceptable use policies for foundation models is likely to suffer from many of the issues digital platforms face in enforcing their content policies [55]. 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 [40, 67, 143]. 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 [148].

6.2 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., blocklists) than for prompts related to producing malware [76]. 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 [165], as would be needed in the case of a model that makes use of multiple other models, as with model merging [27], mixture-of-agents [168], or other systems in which an agent interacts with other models [87]. 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' policies 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 their smartphone, the summarization model in their 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 [116]. This is likely to produce accidental misuse.

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 resources in making their models refuse to generate related content [129]. 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.

6.3 AUPs help shape the foundation model market

Acceptable use policies 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. Many 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 model's outputs [183]. 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 [88].

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 also determine the types of uses of models that are permitted (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.

7. Areas for Future Work

7.1 Collecting data on AUP enforcement

The way in which developers and deployers enforce acceptable use policies for foundation models remains unclear. Collecting data related to enforcement is a key area for future work, as there is little indication that companies will share quantitative data regarding enforcement [14]. 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 depending 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.

7.2. Content moderation on generative AI platforms

Content moderation has been studied much more thoroughly on social media platforms than on AI platforms despite the fact that researchers have access to foundation models but lack access to underlying recommendation systems [100]. Some foundation model developers have adopted content moderation practices quickly, hiring trust and safety teams 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 [63, 130]. 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 [114, 182].

7.3. Regulating acceptable use of foundation models

Governments have attempted to spur self-regulation in the foundation model ecosystem through voluntary codes of conduct [174]. The extent to which governments can go further by forcing developers to block foundation models from generating certain types of content will likely be decided in the courts, much as the ability of social media companies to carry out content moderation has been challenged in the US [155]. Still, China has been somewhat successful in forcing developers to censor the outputs of their foundation models, and many other governments may follow suit [178]. Whether it is feasible for developers to follow such mandates remains an open question, given the ease with which downstream developers can remove safety mitigations.

8. Conclusion

This paper finds that there is significant heterogeneity across acceptable use policies, that they help shape the market for foundation models, and that developers adopt them in order to reduce legal and reputational risks. There is little transparency about acceptable use policies, but this paper takes a first step towards shining a light on why they matter.

Acknowledgments

I thank Ahmed Ahmed, 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

[1] June Ahn, Lauren K. Bivona, and Jeffrey DiScala. 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, 2011. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/meet.2011.14504801044, arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/meet.2011.14504801044, doi:10.1002/meet.2011.14504801044.

[2] Nouf Alfawzan, Markus Christen, Giovanni Spitale, and Nikola Biller-Andorno. Privacy, data sharing, and data security policies of women's mhealth apps: Scoping review and content analysis. *JMIR Mhealth Uhealth*, 10(5):e33735, 2022. URL: https://mhealth.jmir.org/2022/5/e33735, doi:10.2196/33735.

[3] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. URL: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

[4] David Atkinson and Jacob Morrison. A legal risk taxonomy for generative artificial intelligence, 2024. URL: https://arxiv.org/abs/2404.09479, arXiv:2404.09479.

[5] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL: https://arxiv.org/abs/2212.08073, arXiv:2212.08073.

[6] Julia Barnett. 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, August 2023. URL: http://dx.doi.org/10.1145/3600211.3604686, doi:10.1145/3600211.3604686.

[7] Adrien Basdevant, Camille François, Victor Storchan, Kevin Bankston, Ayah Bdeir, Brian Behlendorf, Merouane Debbah, Sayash Kapoor, Yann LeCun, Mark Surman, Helen King-Turvey, Nathan Lambert, Stefano Maffulli, Nik Marda, Govind Shivkumar, and Justine Tunney. Towards a framework for openness in foundation models: Proceedings from the columbia convening on openness in artificial intelligence, 2024. URL: https://arxiv.org/abs/2405.15802, arXiv:2405.15802.

[8] T. Bernier, A. Shah, L. E. Ross, C. H. Logie, and E. Seto. 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, Jun 2021. doi:10.2196/26085.

[9] Katherine E. Beyer. Busting the ghost guns: A technical, statutory, and practical approach to the 3-d printed weapon problem. *Kentucky Law Journal*, 103:433–456, 2014.

[10] Sam Biddle. Openai quietly deletes ban on using chatgpt for "military and warfare". *The Intercept*, January 2024. URL: https://theintercept.com/2024/01/12/open-ai-military-ban-chatgpt/.

[11] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 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, 2022. Association for Computing Machinery. doi:10.1145/3551624.3555290.

[12] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei

Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. arXiv:2108.07258.

[13] Rishi Bommasani, Sayash Kapoor, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Daniel Zhang, Marietje Schaake, Daniel E. Ho, Arvind Narayanan, and Percy Liang. Considerations for governing open foundation models. Technical report, Stanford Institute for Human-Centered AI, December 2023. URL: https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf.

[14] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Betty Xiong, Sayash Kapoor, Nestor Maslej, Arvind Narayanan, and Percy Liang. Foundation model transparency reports, 2024. arXiv:2402.16268.

[15] Rishi Bommasani, Dilara Soylu, Thomas I. Liao, Kathleen A. Creel, and Percy Liang. Ecosystem graphs: The social footprint of foundation models, 2023. arXiv:2303.15772.

[16] Bommasani and Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. The foundation model transparency index v1.1: May 2024, 2024. URL: https://arxiv.org/abs/2407.12929, arXiv:2407.12929.

[17] Bommasani and Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The foundation model transparency index, 2023. arXiv:2310.12941.

[18] Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. The art of saying no: Contextual noncompliance in language models, 2024. URL: https://arxiv.org/abs/2407.12043, arXiv:2407.12043.

[19] Michael Brenes and William D. Hartung. Private finance and the quest to remake modern warfare. Research report, Quincy Institute for Responsible Statecraft, jun 2024. URL: https://quincyinst.org/research/private-finance-and-the-quest-to-remake-modern-warfare/#executive-summary.

[20] Brian Schatz, Ben Ray Luján, Peter Welch, Mark R. Warner, and Angus S. King, Jr. Letter to openai, Jul 2024. URL: https://www.schatz.senate.gov/imo/media/doc/letter_to_openai.pdf.

[21] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. URL: https://arxiv.org/abs/2307.15818, arXiv:2307.15818.

[22] C. Bronstein. Deplatforming sexual speech in the age of fosta/sesta. *Porn Studies*, 8(4):367–380, 2021. doi:10.1080/23268743.2021.1993972.

[23] Sarah Huiyi Cen, Aspen Hopkins, Andrew Ilyas, Aleksander Madry, Isabella Struckman, and Luis Videgaray Caso. Ai supply chains, April 2023. URL: https://ssrn.com/abstract=4789403.

[24] Center for an Informed Public, Digital Forensic Research Lab, Graphika, and Stanford Internet Observatory. The long fuse: Misinformation and the 2020 election, 2021. Stanford Digital Repository: Election Integrity Partnership. v1.3.0. URL: https://purl.stanford.edu/tr171zs0069.

[25] Bilva Chandra, George Awad, Yooyoung Lee, Peter Fontana, Razvan Amironesei, Mark Przybocki, Kamie Roberts, Elham Tabassi, Mat Heyman, and Jesse Dunietz. Reducing risks posed by synthetic content, April 2024. URL: https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf.

[26] Nicole Chi, Emma Lurie, and Deirdre K. Mulligan. Reconfiguring diversity and inclusion for ai ethics. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 447–457,

New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3461702.3462622.

[27] Leshem Choshen, Elad Venezian, Noam Slonim, and Yoav Katz. Fusing finetuned models for better pretraining, 2022. URL: https://arxiv.org/abs/2204.03044, arXiv:2204.03044.

[28] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL: https://arxiv.org/abs/1706.03741, arXiv:1706.03741.

[29] Cohere, OpenAI, and AI21 Labs. Best practices for deploying language models, Jul 2022. URL: https://cdn.openai.com/papers/joint-recommendation-for-language-model-deployment.pdf.

[30] Samantha Cole. Riley reid on ai: 'i don't want porn to get left behind', Oct 2023. URL: https://www.404media.co/riley-reid-clona-ai-chatbot-virtual-companion/.

[31] Competition and Markets Authority. Ai foundation models initial report. Report, UK Competition & Markets Authority, Sep 2023. URL: https://assets.publishing.service.gov.uk/media/650449e86771b90014fdab4c/Full_Non-Confidential_Report_PDFA.pdf.

[32] Competition and Markets Authority. Ai foundation models: Technical update report. Technical report, UK Competition & Markets Authority, Apr 2024. URL: https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf.

[33] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. Behavioral use licensing for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. ACM, June 2022. URL: http://dx.doi.org/10.1145/3531146.3533143, doi:10.1145/3531146.3533143.

[34] A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 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, page 864–876, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3531146.3533150.

[35] Ned Cooper and Alexandra Zafiroglu. 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, 2024. Association for Computing Machinery. doi:10.1145/3613904.3642775.

[36] Cyberspace Administration of China. Interim measures for the management of generative artificial intelligence services. https://www.chinalawtranslate.com/en/generative-ai-interim/, July 2023. Accessed: 2024-05-14.

[37] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93, January 2024. URL: http://dx.doi.org/10.1093/jla/laae003, doi:10.1093/jla/laae003.

[38] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 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, 2023. Association for Computing Machinery. doi:10.1145/3617694.3623261.

[39] Neil Doherty, Leonidas Anastasakis, and Heather Fulford. 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, 06 2011. doi:10.1016/j.ijinfomgt.2010.06.001.

[40] J. Donovan, E. Dreyfuss, and B. Friedberg. *Meme Wars: The Untold Story of the Online Battles Upending Democracy in America*. Bloomsbury Publishing, 2022. URL: https://books.google.com/books?id=04l3EAAAQBAJ.

[41] Evelyn Douek. Content moderation as systems thinking. *Harvard Law Review*, 2022. URL: https://ssrn.com/abstract=4005326, doi:10.2139/ssrn.4005326.

[42] Evelyn Douek. The meta oversight board and the empty promise of legitimacy. *Harvard Journal of Law & Technology*, 37, 2024. doi:10.2139/ssrn.4565180.

[43] Kate Downing. Ai licensing can't balance "open" with "responsible", Jul 2023. URL: https://katedowninglaw.com/2023/07/13/ai-licensing-cant-balance-open-with-responsible/.

[44] Kate Downing. Choose your own adventure: The eu ai act and openish ai, February 2024. URL: https://katedowninglaw.com/2023/07/13/ai-licensing-cant-balance-open-with-responsible/.

[45] Francisco Eiras, Aleksandar Petrov, Bertie Vidgen, Christian Schroeder, Fabio Pizzati, Katherine Elkins, Supratik Mukhopadhyay, Adel Bibi, Aaron Purewal, Csaba Botos, Fabro Steibel, Fazel Keshtkar, Fazl Barez, Genevieve Smith, Gianluca Guadagni, Jon

Chun, Jordi Cabot, Joseph Imperial, Juan Arturo Nolazco, Lori Landay, Matthew Jackson, Phillip H. S. Torr, Trevor Darrell, Yong Lee, and Jakob Foerster. Risks and opportunities of open-source generative ai, 2024. URL: https://arxiv.org/abs/2405.08597, arXiv:2405.08597.

[46] Satu Elo and Helvi Kyngäs. The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1):107–115, 2008. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2648.2007.04569.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2648.2007.04569.x, doi:10.1111/j.1365-2648.2007.04569.x.

[47] Jerel M. Ezell, Babatunde Patrick Ajayi, Tapan Parikh, Kyle Miller, Alex Rains, and David Scales. Drug use and artificial intelligence: Weighing concerns and possibilities for prevention. *American Journal of Preventive Medicine*, 66(3):568–572, 2024. URL: https://www.sciencedirect.com/science/article/pii/S0749379723004841, doi:10.1016/j.amepre.2023.11.024.

[48] Steven Feldstein. *The global expansion of AI surveillance*. Carnegie Endowment for International Peace Washington, DC, 2019.

[49] Grant Fergusson, Caitriona Fitzgerald, Chris Frascella, Megan Iorio, Tom McBrien, Calli Schroeder, Ben Winters, and Enid Zhou. Generating harms: Generative ai's impact & paths forward. Technical report, Electronic Privacy Information Center, 2023. URL: https://epic.org/documents/generating-harms-generative-ais-impact-paths-forward/.

[50] Thomas Ferretti. An institutionalist approach to ai ethics: Justifying the priority of government regulation over self-regulation. *Moral Philosophy and Politics*, 9(2):239–265, 2022. doi:10.1515/mopp-2020-0056.

[51] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication*, January 2020. http://dx.doi.org/10.2139/ssrn.3518482.

[52] Luciano Floridi. Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, 32:185–193, 2019. doi:10.1007/s13347-019-00354-x.

[53] Aakash Gautam. Reconfiguring participatory design to resist ai realism. *arXiv preprint arXiv:2406.03245*, 2024. Presented at Participatory Design Conference 2024. URL: https://doi.org/10.48550/arXiv.2406.03245, arXiv:2406.03245.

[54] Yinuo Geng. Comparing" deepfake" regulatory regimes in the united states, the european union, and china. *Geo. L. Tech. Rev.*, 7:157, 2023.

[55] T. Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press, 2018. URL: https://books.google.com/books?id=-RteDwAAQBAJ.

[56] GitHub. Github acceptable use policies, 2024. Accessed: July 2024. URL: https://docs.github.com/en/site-policy/acceptable-use-policies/github-acceptable-use-policies.

[57] Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. How persuasive is AI-generated propaganda? *PNAS Nexus*, 3(2):pgae034, 02 2024. arXiv:https://academic.oup.com/pnasnexus/article-pdf/3/2/pgae034/56712546/pgae034.pdf, doi:10.1093/pnasnexus/pgae034.

[58] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. Generative language models and automated influence operations: Emerging threats and potential mitigations, 2023. URL: https://arxiv.org/abs/2301.04246, arXiv:2301.04246.

[59] Google. Policy guidelines for the gemini app, 2024. URL: https://gemini.google/policy-guidelines/.

[60] Robert Gorwa and Michael Veale. Moderating model marketplaces: Platform governance puzzles for ai intermediaries, 2024. arXiv:2311.12573.

[61] Pauline Gourlet, Donato Ricci, and Maxime Crépel. Reclaiming artificial intelligence accounts: A plea for a participatory turn in artificial intelligence inquiries. *Big Data & Society*, 11(2):20539517241248093, 2024. arXiv:https://doi.org/10.1177/20539517241248093, doi:10.1177/20539517241248093.

[62] Declan Grabb, Max Lamparth, and Nina Vasan. Risks from language models for automated mental healthcare: Ethics and structure for implementation. *medRxiv*, 2024. URL: https://www.medrxiv.org/content/early/2024/04/08/2024.04.07.24305462, arXiv:https://www.medrxiv.org/content/early/2024/04/08/2024.04.07.24305462.full.pdf, doi:10.1101/2024.04.07.24305462.

[63] M.L. Gray and S. Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt, 2019. URL: https://books.google.com/books?id=u10-uQEACAAJ.

[64] Nick Gregorio, Janahan Mathanamohan, Qusay H. Mahmoud, and May AlTaei. Hacking in the cloud. *Internet Technology Letters*, 2(1):e84, 2019. URL:

https://onlinelibrary.wiley.com/doi/abs/10.1002/itl2.84, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/itl2.84, doi:10.1002/itl2.84.

[65] Philipp Hacker. Ai regulation in europe: From the ai act to future regulatory challenges, 2023. URL: https://arxiv.org/abs/2310.04072, arXiv:2310.04072.

[66] Philipp Hacker, Johann Cordes, and Janina Rochon. 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, 2024.

[67] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 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, 2021.

[68] Vaughn Hamilton, Hanna Barakat, and Elissa M. Redmiles. Risk, resilience and reward: Impacts of shifting to digital sex work. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), nov 2022. doi:10.1145/3555650.

[69] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. Foundation models and fair use, 2023. arXiv:2303.15715.

[70] Peter Henderson, Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, and Prateek Mittal. Safety risks from customizing foundation models via fine-tuning. Policy brief, Stanford Institute for Human-Centered AI, January 2024. URL: https://hai.stanford.edu/sites/default/files/2024-01/Policy-Brief-Safety-Risks-Customizing-Foundation-Models-Fine-Tuning.pdf.

[71] Mia Hoffmann and Heather Frase. Adding structure to ai harm: An introduction to cset's ai harm framework. Technical report, Center for Security and Emerging Technology, July 2023. doi:10.51593/20230022.

[72] Krystal Hu, Greg Bensinger, and Jody Godoy. Exclusive: Ftc seeking details on amazon deal with ai startup adept, source says. *Reuters*, Jul 2024. Accessed [Insert Access Date]. URL: https://www.reuters.com/technology/ftc-seeking-details-amazon-deal-with-ai-startup-adept-source-says-2024-07-16/.

[73] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective constitutional ai: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1395–1417, New York, NY, USA, 2024. Association for Computing Machinery. doi:10.1145/3630106.3658979.

[74] Hugging Face. Content policy, August 2023. URL: https://huggingface.co/content-guidelines.

[75] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. URL: https://arxiv.org/abs/2312.06674, arXiv:2312.06674.

[76] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):1–33, 2018.

[77] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, September 2019. doi:10.1038/s42256-019-0088-2.

[78] A. Jones. *Camming: Money, Power, and Pleasure in the Sex Work Industry*. NYU Press, 2020. URL: https://books.google.com/books?id=30SODwAAQBAJ.

[79] Barbara M. Jones. 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, 2015. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/bul2.2015.1720420113, arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/bul2.2015.1720420113, doi:10.1002/bul2.2015.1720420113.

[80] Nektaria Kaloudi and Jingyue Li. The ai-based cyber threat landscape: A survey. *ACM Comput. Surv.*, 53(1), feb 2020. doi:10.1145/3372823.

[81] Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, Rumman Chowdhury, Alex Engler, Peter Henderson, Yacine Jernite, Seth Lazar, Stefano Maffulli, Alondra Nelson, Joelle Pineau, Aviya Skowron, Dawn Song, Victor Storchan, Daniel Zhang, Daniel E. Ho, Percy Liang, and Arvind Narayanan. On the societal impact of open foundation models, 2024. arXiv:2403.07918.

[82] Rishabh Kaushal, Jacob van de Kerkhof, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. Automated transparency: A legal and empirical anal-

ysis of the digital services act transparency database, 2024. arXiv:2404.02894.

[83] Imrul Kayes and Adriana Iamnitchi. Privacy and security in online social networks: A survey. *Online Social Networks and Media*, 3:1–21, 2017.

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

[85] Paul Keller and Nicolò Bonato. Growth of responsible AI licensing. Analysis of license use for ML models published on. *Open Future*, feb 7 2023. URL: https://openfuture.pubpub.org/pub/growth-of-responsible-ai-licensing.

[86] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model, 2024. URL: https://arxiv.org/abs/2406.09246, arXiv:2406.09246.

[87] Shiyang Lai, Yujin Potter, Junsol Kim, Richard Zhuang, Dawn Song, and James Evans. Position: Evolving AI collectives enhance human diversity and enable self-regulation. In *Forty-first International Conference on Machine Learning*, 2024. URL: https://openreview.net/forum?id=u6PeRHEsjL.

[88] Nathan Lambert. Llama 3.1 405b, meta's ai strategy, and the new, open frontier model ecosystem. *Interconnects*, Jul 2024. URL: https://www.interconnects.ai/p/llama-405b-open-frontier-model.

[89] Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. The history and risks of reinforcement learning and human feedback, 2023. arXiv:2310.13595.

[90] Katherine Lee, A. Feder Cooper, and James Grimmelmann. Talkin' 'bout ai generation: Copyright and the generative-ai supply chain, 2024. arXiv:2309.08133.

[91] V. Lehdonvirta. *Cloud Empires: How Digital Platforms Are Overtaking the State and How We Can Regain Control*. MIT Press, 2022. URL: https://books.google.com/books?id=bc9UEAAAQBAJ.

[92] Mark A. Lemley, Peter Henderson, and Tatsunori Hashimoto. Where's the liability in harmful ai speech?, August 2023. http://dx.doi.org/10.2139/ssrn.4531029.

[93] Han Li, Jie Zhang, and Rathindra Sarathy. Understanding compliance with internet use policy from the perspective of rational choice theory. *Decision Support Systems*, 48(4):635–645, 2010. URL: https://www.sciencedirect.com/science/article/pii/S0167923609002619, doi:10.1016/j.dss.2009.12.005.

[94] Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. What's documented in ai? systematic analysis of 32k ai model cards, 2024. URL: https://arxiv.org/abs/2402.05160, arXiv:2402.05160.

[95] Shayne Longpre, Stella Biderman, Alon Albalak, Hailey Schoelkopf, Daniel McDuff, Sayash Kapoor, Kevin Klyman, Kyle Lo, Gabriel Ilharco, Nay San, Maribeth Rauh, Aviya Skowron, Bertie Vidgen, Laura Weidinger, Arvind Narayanan, Victor Sanh, David Adelani, Percy Liang, Rishi Bommasani, Peter Henderson, Sasha Luccioni, Yacine Jernite, and Luca Soldaini. The responsible foundation model development cheatsheet: A review of tools & resources, 2024. URL: https://arxiv.org/abs/2406.16746, arXiv:2406.16746.

[96] Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, Yi Zeng, Weiyan Shi, Xianjun Yang, Reid Southen, Alexander Robey, Patrick Chao, Diyi Yang, Ruoxi Jia, Daniel Kang, Sandy Pentland, Arvind Narayanan, Percy Liang, and Peter Henderson. A safe harbor for ai evaluation and red teaming, 2024. arXiv:2403.04893.

[97] Seán Looney. Content moderation through removal of service: Content delivery networks and extremist websites. *Policy & Internet*, 15(4):544–558, 2023. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.370, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/poi3.370, doi:10.1002/poi3.370.

[98] Alexandra Loverock, Tyler Marshall, Dylan Viste, Fahad Safi, Will Rioux, Navid Sedaghat, Megan Kennedy, and S. Monty Ghosh. Electronic harm reduction interventions for drug overdose monitoring and prevention: A scoping review. *Drug and Alcohol Dependence*, 250:110878, 2023. URL: https://www.sciencedirect.com/science/article/pii/S037687162301116X, doi:10.1016/j.drugalcdep.2023.110878.

[99] Michael J Madison. Reconstructing the software license. *Loy. U. Chi. Lj*, 35:275, 2003.

[100] Yaaseen Mahomed, Charlie M. Crawford, Sanjana Gautam, Sorelle A. Friedler, and Danaë Metaxa. 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, page 660–686, New York, NY, USA, 2024. Association for Computing Machinery. doi:10.1145/3630106.3658932.

[101] Nahema Marchal, Rachel Xu, Rasmi Elasmar, Iason

Gabriel, Beth Goldberg, and William Isaac. Generative ai misuse: A taxonomy of tactics and insights from real-world data, 2024. URL: https://arxiv.org/abs/2406.13843, arXiv:2406.13843.

[102] G Alan Marlatt. Harm reduction: Come as you are. *Addictive behaviors*, 21(6):779–788, 1996.

[103] Glenn Ellingson Matt Motyl. The unbearably high cost of cutting trust & safety corners, 2024. URL: https://www.techpolicy.press/the-unbearably-high-cost-of-cutting-trust-safety-corners/.

[104] Natalie Maus, Patrick Chao, Eric Wong, and Jacob R Gardner. Black box adversarial prompting for foundation models. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

[105] Philipp Mayring, Angelika Bikner-Ahsbahs, Christine Knipping, and Norma Presmeg. *Qualitative Content Analysis: Theoretical Background and Procedures*, pages 365–380. Springer Netherlands, Dordrecht, 2015. doi:10.1007/978-94-017-9181-6_13.

[106] Daniel McDuff, Tim Korjakow, Scott Cambo, Jesse Josua Benjamin, Jenny Lee, Yacine Jernite, Carlos Muñoz Ferrandis, Aaron Gokaslan, Alek Tarkowski, Joseph Lindley, A. Feder Cooper, and Danish Contractor. On the standardization of behavioral use clauses and their adoption for responsible licensing of ai, 2024. arXiv:2402.05979.

[107] D. McMenemy, University of Strathclyde. Department of Computer, and Information Sciences. Public library digital services: Emergent issues of access and acceptable use, 2019. URL: https://books.google.com/books?id=RDmO0AEACAAJ.

[108] Michelle M. Mello and Neel Guha. Understanding liability risk from using health care artificial intelligence tools. *New England Journal of Medicine*, 390(3):271–278, 2024. URL: https://www.nejm.org/doi/full/10.1056/NEJMhle2308901, arXiv:https://www.nejm.org/doi/pdf/10.1056/NEJMhle2308901, doi:10.1056/NEJMhle2308901.

[109] Rachel Metz and Brody Ford. Adobe's 'ethical' firefly ai was trained on midjourney images. *Bloomberg*, April 2024. URL: https://www.bloomberg.com/news/articles/2024-04-12/adobe-s-ai-firefly-used-ai-generated-images-from-rivals-for-training.

[110] Mary Minow, Tomas A. Lipinski, Gretchen McCord, et al. *The Library's Legal Answers for Makerspaces*. ALA Editions, 2016. eBook.

[111] Margaret Mitchell. The pillars of a rights-based approach to ai development, 2023. URL: https://www.techpolicy.press/the-pillars-of-a-rightsbased-approach-to-ai-development/.

[112] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3287560.3287596.

[113] Christopher A. Mouton, Caleb Lucas, and Ella Guest. *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study*. RAND Corporation, Santa Monica, CA, 2024. doi:10.7249/RRA2977-2.

[114] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models, 2020. URL: https://arxiv.org/abs/2010.00133, arXiv:2010.00133.

[115] Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. Conceptualizing ai literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2:100041, 2021. URL: https://www.sciencedirect.com/science/article/pii/S2666920X21000357, doi:10.1016/j.caeai.2021.100041.

[116] Jonathan A. Obar and Anne Oeldorf-Hirsch. The biggest lie on the internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147, 2020. arXiv:https://doi.org/10.1080/1369118X.2018.1486870, doi:10.1080/1369118X.2018.1486870.

[117] W. Ian O'Byrne. *Acceptable Use Policies*, pages 1–6. John Wiley & Sons, Ltd, 2019. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118978238.ieml0001, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118978238.ieml0001, doi:10.1002/9781118978238.ieml0001.

[118] National Technical Committee 260 on Cybersecurity of Standardization Administration of China (SAC/TC260). Basic safety requirements for generative artificial intelligence services, April 2024. Translated by the Center for Security and Emerging Technology. URL: https://cset.georgetown.edu/publication/china-safety-requirements-for-generative-ai-final/.

[119] Open Source Initiative. The Open Source AI Definition – draft v. 0.0.8, 2024. Accessed July 2024. URL: https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-8.

[120] OpenAI. Model spec, May 2024. URL: https://cdn.openai.com/spec/model-spec-2024-05-08.html.

[121] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. arXiv:2303.08774.

[122] Aviv Ovadya. Reimagining democracy for ai. *Journal of Democracy*, 34(4):162–170, Oct 2023. URL: https://www.journalofdemocracy.org/articles/reimagining-democracy-for-ai/.

[123] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 2016 ACM International Conference on Supporting Group Work*, GROUP '16, page 369–374, New York, NY, USA, 2016. Association for Computing Machinery. doi:10.1145/2957276.2957297.

[124] Riana Pfefferkorn. Addressing computer-generated child sex abuse imagery: Legal framework and policy implications. *Lawfare*, Feb 2024. Accessed [Insert Access Date]. URL: https://www.lawfaremedia.org/article/addressing-computer-generated-child-sex-abuse-imagery-legal-framework-and-policy-implications.

[125] Giada Pistilli, Carlos Muñoz Ferrandis, Yacine Jernite, and Margaret Mitchell. 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, page 343–354, New

York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3593013.3594002.

[126] Priyanshu Priya, Mauajama Firdaus, and Asif Ekbal. Computational politeness in natural language processing: A survey. *ACM Computing Surveys*, 56(9):1–42, May 2024. URL: http://dx.doi.org/10.1145/3654660, doi:10.1145/3654660.

[127] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to!, 2023. arXiv:2310.03693.

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

[129] Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Alexandra Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel J. Kochenderfer, and Robert Trager. Open problems in technical ai governance, 2024. URL: https://arxiv.org/abs/2407.14981, arXiv:2407.14981.

[130] S.T. Roberts. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, 2019. URL: https://books.google.com/books?id=uiCbDwAAQBAJ.

[131] Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

[132] Elaine Robinson. The panoptic principle: privacy and surveillance in the public library as evidenced in the acceptable use policy. Thesis, University of Strathclyde, 2019. Accessed: 2021-07-01. URL: http://localhost/files/gq67jr277.

[133] Elaine Robinson and David McMenemy. '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, 2020. arXiv:https://doi.org/10.1177/0961000619871598, doi:10.1177/0961000619871598.

[134] A.B. Ruighaver, S.B. Maynard, and M. Warren. Ethical decision making: Improving the quality of acceptable use policies. *Computers & Security*, 29(7):731–736, 2010. URL: https://www.sciencedirect.com/science/article/pii/S0167404810000386, doi:10.1016/j.cose.2010.05.004.

[135] Teela Sanders, Jane Scoular, Rosie Campbell, Jane Pitcher, and Stewart Cunningham. *Internet sex work: Beyond the gaze*. Springer, 2018.

[136] Johannes Schneider, Arianna Casanova Flores, and Anne-Catherine Kranz. Exploring human-llm conversations: Mental models and the originator of toxicity, 2024. URL: https://arxiv.org/abs/2407.05977, arXiv:2407.05977.

[137] Elizabeth Seger, Aviv Ovadya, Ben Garfinkel, Divya Siddarth, and Allan Dafoe. Democratising ai: Multiple meanings, goals, and methods, 2023. URL: https://arxiv.org/abs/2303.12642, arXiv:2303.12642.

[138] Rusheb Shah, Quentin Feuillade Montixi, Soroush Pour, Arush Tagade, and Javier Rando. Scalable and transferable black-box jailbreaks for language models via persona modulation. In *Socially Responsible Language Modelling Research*, 2023.

[139] Megan Shahi, Adam Conner, and Nicole Alvarez. Generative ai should be developed and deployed responsibly at every level for everyone, February 1 2024. URL: https://www.americanprogress.org/article/generative-ai-should-be-developed-and-deployed-responsibly-at-every-level-for-everyone/.

[140] Gemma Sharp, John Torous, and Madeline L. West. Ethical challenges in ai approaches to eating disorders. *Journal of Medical Internet Research*, 25:e50696, August 2023. ©Gemma Sharp, John Torous, Madeline L. West. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 14.08.2023. URL: https://mhealth.jmir.org/2022/5/e33735, doi:10.2196/50696.

[141] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction, 2023. arXiv:2210.05791.

[142] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*, 2023.

[143] Eugenia Siapera and Paloma Viejo-Otero. Governing hate: Facebook and digital racism. *Television & New Media*, 22(2):112–130, 2021.

[144] Keng Siau, Fiona Fui-Hoon Nah, and Limei Teng. Acceptable internet use policy. *Commun. ACM*, 45(1):75–79, jan 2002. doi:10.1145/502269.502302.

[145] Riley Simmons-Edler, Ryan Badman, Shayne Longpre, and Kanaka Rajan. Ai-powered autonomous weapons risk geopolitical instability and threaten ai

research, 2024. URL: https://arxiv.org/abs/2405.01859, arXiv:2405.01859.

[146] Zachary Small. Black artists say a.i. shows bias, with algorithms erasing their history. *The New York Times*, July 2023. URL: https://www.nytimes.com/2023/07/04/arts/design/black-artists-bias-ai.html.

[147] Irene Solaiman. The gradient of generative ai release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 111–122, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3593013.3593981.

[148] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé III au2, Jesse Dodge, Isabella Duan, Ellie Evans, Felix Friedrich, Avijit Ghosh, Usman Gohar, Sara Hooker, Yacine Jernite, Ria Kalluri, Alberto Lusoli, Alina Leidinger, Michelle Lin, Xiuzhu Lin, Sasha Luccioni, Jennifer Mickel, Margaret Mitchell, Jessica Newman, Anaelia Ovalle, Marie-Therese Png, Shubham Singh, Andrew Strait, Lukas Struppek, and Arjun Subramonian. Evaluating the social impact of generative ai systems in systems and society, 2024. URL: https://arxiv.org/abs/2306.05949, arXiv:2306.05949.

[149] Madhulika Srikumar, Jiyoo Chang, and Kasia Chmielinski. Risk mitigation strategies for the open foundation model value chain, July 11 2024. URL: https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/.

[150] Zahra Stardust. Safe for work: Feminist porn, corporate regulation and community standards. In Catherine Dale and Rosemary Overell, editors, *Orienting Feminism: Media, Activism and Cultural Representation*, pages 155–179. Springer International Publishing, Cham, 2018. URL: https://link.springer.com/chapter/10.1007/978-3-319-70660-3_9.

[151] Farley Stewart. Internet acceptable use policies: Navigating the management, legal, and technical issues. *Information Systems Security*, 9(3):1–7, 2000. arXiv:https://doi.org/10.1201/1086/43310.9.3.20000708/31360.6, doi:10.1201/1086/43310.9.3.20000708/31360.6.

[152] Angelika Strohmayer, Jenn Clamen, and Mary Laing. Technologies for social justice: Lessons from sex workers on the front lines. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–14, 2019.

[153] Substance Abuse and Mental Health Services Administration. Harm reduction framework. Technical report, Center for Substance Abuse Prevention, Substance Abuse and Mental Health Services Administration, 2023. URL: https://www.samhsa.gov/sites/default/files/harm-reduction-framework.pdf.

[154] Lucy Suchman. Algorithmic warfare and the reinvention of accuracy. *Critical Studies on Security*, 8(2):175–187, 2020. arXiv:https://doi.org/10.1080/21624887.2020.1760587, doi:10.1080/21624887.2020.1760587.

[155] Supreme Court of the United States. Netchoice, llc v. paxton. United States Supreme Court, Jul 2024. Docket No. 22-555, 598 U.S. (2024). URL: https://www.supremecourt.gov/opinions/23pdf/22-555_h3ci.pdf.

[156] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1609–1621, New York, NY, USA, 2024. Association for Computing Machinery. doi:10.1145/3630106.3658992.

[157] Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions, 2023. URL: https://arxiv.org/abs/2312.03689, arXiv:2312.03689.

[158] David Thiel, Melissa Stroebel, and Rebecca Portnoff. Generative ml and csam: Implications and mitigations. Report, Stanford Internet Observatory, Jun 2023. URL: https://purl.stanford.edu/jv206yg3793.

[159] Thorn. Safety by design for generative ai: Preventing child sexual abuse, 2024. URL: https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf.

[160] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[161] Jon Truby, Rafael Dean Brown, Imad Antoine Ibrahim, and Oriol Caudevilla Parellada. A sandbox approach to regulating high-risk artificial intelligence applications. *European Journal of Risk Regulation*, 13(2):270–294, 2022.

[162] European Union. 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, 2024. Accessed: 2024-05-14. URL: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf.

[163] United States District Court for the District of Columbia. United states et al. v. google llc, Aug 2024. Case No. 20-cv-3010 (APM), Memorandum Opinion. URL: https://storage.courtlistener.com/recap/gov.uscourts.dcd.223205/gov.uscourts.dcd.223205.1033.0_1.pdf.

[164] Margrethe Vestager, Sarah Cardell, Jonathan Kanter, and Lina M. Khan. Joint statement on competition in generative ai foundation models and ai products, Jul 2024. URL: https://www.ftc.gov/system/files/ftc_gov/pdf/ai-joint-statement.pdf.

[165] Luis Villa. Evaluating the rail license family, November 2022. URL: https://blog.tidelift.com/evaluating-the-rail-license-family.

[166] Sandra Wachter and Brent Mittelstadt. A right to reasonable inferences: re-thinking data protection law in the age of big data and ai. *Colum. Bus. L. Rev.*, page 494, 2019.

[167] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, 2024. arXiv:2306.11698.

[168] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities, 2024. URL: https://arxiv.org/abs/2406.04692, arXiv:2406.04692.

[169] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

[170] Laura Weidinger, John Mellor, Bernat Guillen Pegueroles, Nahema Marchal, Ravin Kumar, Kristian Lum, Canfer Akbulut, Mark Diaz, Stevie Bergman, Mikel Rodriguez, Verena Rieser, and William Isaac. Star: Sociotechnical approach to red teaming language models, 2024. URL: https://arxiv.org/abs/2406.11757, arXiv:2406.11757.

[171] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. Sociotechnical safety evaluation of generative ai systems, 2023. arXiv:2310.11986.

[172] Jake Weidman and Jens Grossklags. 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)*, page Research Papers, Stockholm & Uppsala, Sweden, 2019. URL: https://aisel.aisnet.org/ecis2019_rp/99.

[173] Matt White, Ibrahim Haddad, Cailean Osborne, Xiao-Yang Liu Yanglet, Ahmed Abdelmonsef, and Sachin Varghese. The model openness framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence, 2024. URL: https://arxiv.org/abs/2403.13784, arXiv:2403.13784.

[174] White House. Voluntary ai commitments, July 2023. Accessed: 2024-05-14. URL: https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf.

[175] David Gray Widder. Epistemic power in ai ethics labor: Legitimizing located complaints. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1295–1304, New York, NY, USA, 2024. Association for Computing Machinery. doi:10.1145/3630106.3658973.

[176] David Gray Widder, Sarah West, and Meredith Whittaker. Open (for business): Big tech, concentrated power, and the political economy of open ai. *SSRN*, 2023. URL: https://ssrn.com/abstract=4543807, doi:10.2139/ssrn.4543807.

[177] Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023.

[178] Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. Air-bench 2024: A safety benchmark based on risk categories from regulations and policies, 2024. URL: https://arxiv.org/abs/2407.17436, arXiv:2407.17436.

[179] Zeng and Klyman, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. Ai risk categorization decoded (air 2024): From government regulations to corporate policies, 2024. URL: https://arxiv.org/abs/2406.17864, arXiv:2406.17864.

[180] Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. Removing RLHF Protections in GPT-4 via Fine-Tuning. *arXiv preprint arXiv:2311.05553*, 2023.

[181] Angela Huyue Zhang. The promise and perils of china's regulation of artificial intelligence. *University of Hong Kong Faculty of Law Research Paper*, 2024(02), 2024. http://dx.doi.org/10.2139/ssrn.4708676. URL: https://ssrn.com/abstract=4708676.

[182] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods, 2018. URL: https://arxiv.org/abs/1804.06876, arXiv:1804.06876.

[183] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild, 2024. URL: https://arxiv.org/abs/2405.01470, arXiv:2405.01470.

[184] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.