Position: On the Societal Impact of Open Foundation Models

Sayash Kapoor [*1] **Rishi Bommasani** [*2]
Kevin Klyman [2] **Shayne Longpre** [3] **Ashwin Ramaswami** [4] **Peter Cihon** [5] **Aspen Hopkins** [3]
Kevin Bankston [6 4] **Stella Biderman** [7] **Miranda Bogen** [6] **Rumman Chowdhury** [8] **Alex Engler** [9]
Peter Henderson [1] **Yacine Jernite** [10] **Seth Lazar** [11] **Stefano Maffulli** [12] **Alondra Nelson** [13] **Joelle Pineau** [14]
Aviya Skowron [7] **Dawn Song** [15] **Victor Storchan** [16] **Daniel Zhang** [2]
Daniel E. Ho [2] **Percy Liang** [2] **Arvind Narayanan** [1]

Abstract

Foundation models are powerful technologies: how they are released publicly directly shapes their societal impact. In this position paper, we focus on *open* foundation models, defined here as those with broadly available model weights (e.g. Llama 3, Stable Diffusion XL). We identify five distinctive properties of open foundation models (e.g. greater customizability, poor monitoring) that mediate their benefits and risks. Open foundation models present significant benefits, with some caveats, that span innovation, competition, the distribution of decision-making power, and transparency. To understand their risks of misuse, we design a risk assessment framework for analyzing their *marginal risk*. Across several misuse vectors (e.g. cyberattacks, bioweapons), we find that current research is insufficient to effectively characterize the marginal risk of open foundation models relative to pre-existing technologies. The framework helps explain why the marginal risk is low in some cases, clarifies disagreements about misuse risks by revealing that past work has focused on different subsets of the framework with different assumptions, and articulates a way forward for more constructive debate. Overall, our work supports a more grounded assessment of the societal impact of open foundation models by outlining what research is needed to empirically validate their theoretical benefits and risks.

[*]Equal contribution [1]Princeton University [2]Stanford University [3]Massachusetts Institute of Technology [4]Georgetown University [5]GitHub [6]Center for Democracy and Technology [7]Eleuther AI [8]Humane Intelligence [9]Brookings Institution [10]Hugging Face [11]Australian National University [12]Open Source Initiative [13]Institute for Advanced Study [14]Meta [15]University of California, Berkeley [16]Mozilla AI. Correspondence to: Sayash Kapoor <sayashk@princeton.edu>, Rishi Bommasani <nlprishi@stanford.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Foundation models (Bommasani et al., 2021) are the centerpiece of the modern AI ecosystem, catalyzing a frenetic pace of technological development, deployment, and adoption that brings with it controversy, scrutiny, and public attention. *Open foundation models*[1] like BERT, CLIP, Whisper, BLOOM, Pythia, Llama 2, Falcon, Stable Diffusion, Mistral, OLMo, Aya, Gemma, and Phi-3 play an important role in this ecosystem. These models allow greater customization and deeper inspection of how they operate, giving developers greater choice in selecting foundation models. However, they may also increase risk, especially given broader adoption, which has prompted pushback, especially around risks relating to biosecurity, cybersecurity, and disinformation. How to release foundation models is a central debate today, often described as open vs. closed.

Simultaneously, policymakers are confronting how to govern open foundation models. In the United States, the recent Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence mandates that the Department of Commerce prepare a report for the President on the benefits and risks of open foundation models (Executive Office of the President, 2023). In the European Union, open foundation models are partially exempt from obligations under the recently-negotiated AI Act. And consideration of widely available model weights is a stated priority of the UK's AI Safety Institute (UK AISI, 2023).

Given disagreement within the AI community and uncertainty on appropriate AI policy (§2), **in this paper, we clarify the benefits and risks of open foundation models**.

[1]We define *open foundation models* as foundation models with widely available model weights (see Executive Office of the President, 2023; National Telecommunications and Information Administration, 2024).

We decompose the analysis of the societal impact of open foundation models into two steps. First, we articulate five distinctive properties of open foundation models (§3). Open foundation models are marked by broader access, greater customizability, the potential for local inference, an inability to rescind model access once released, and weaker monitoring.

Second, we outline how these distinctive properties lead to specific benefits and risks of open foundation models. The benefits we identify are distributing decision-making power, reducing market concentration, increasing innovation, accelerating science, and enabling transparency (§4). We highlight considerations that may temper these benefits in practice (e.g. model weights are sufficient for some forms of science, but access to training data is necessary for others and is not guaranteed by release of weights).

Turning to risks, we present a framework for conceptualizing the *marginal risk* of open foundation models: that is, the extent to which these models increase societal risk by intentional misuse beyond closed foundation models or pre-existing technologies, such as web search on the internet (§5). Surveying seven common misuse vectors described for open foundation models (e.g. disinformation, biosecurity, cybersecurity, non-consensual intimate imagery, scams), we find that past studies do not clearly assess the marginal risk in most cases.

Our framework helps explain why the marginal risk is low in some cases where we already have evidence from past waves of digital technology (such as the use of foundation models for automated vulnerability detection in cybersecurity). It also helps retrospectively explain why the research on the dangers of open foundation models has been so contentious—past studies implicitly analyze risks for different subsets of our framework. The framework provides a way to have a more productive debate going forward, by outlining the necessary components of a complete analysis of the misuse risk of open foundation models. Namely, while the current evidence for marginal risk is weak for several misuse vectors, we encourage more empirically grounded work to assess the marginal risk, recognizing the nature of this risk will evolve as model capabilities and societal defenses evolve.

By clearly articulating the benefits and risks of open foundation models, including where current empirical evidence is lacking, we ground ongoing discourse and policymaking. Specifically, we use our analysis to direct recommendations at AI developers, researchers investigating the risks of AI, competition regulators, and policymakers (§6). Action from these stakeholders can further clarify the societal impact of open foundation models and, thereby, enhance our ability to reap their benefits while mitigating risks.

2. Background

The release landscape for foundation models is complex (Sastry, 2021; Liang et al., 2022a; Solaiman, 2023; Basdevant et al., 2024; Nelson et al., 2024). In particular, several *assets* exist (e.g. the model, data, code): for each asset, there is the matter of *who* can access the asset (e.g. user restrictions like requiring that the user be 18 or older) and for *what* purposes (e.g. use restrictions that prohibit usage for competing against the model developer).[2] Further, the degree of access may *change over time* (e.g. staged release to broaden access, deprecation to reduce access).

In this paper, we consider a reductive, but useful, dichotomy between open and closed foundation models to facilitate analysis. We define *open foundation models* as foundation models with widely available model weights. (For simplicity, we refer to any non-open foundation model as *closed*.) In particular, with respect to the dimensions of release we describe, this means an open foundation model (i) must provide weights-level access, (ii) need not be accompanied by the open release of any other assets (e.g. code, data, or compute), (iii) must be widely available, though some restrictions on users (e.g. based on age) may apply, (iv) need not be released in stages, and (v) may have use restrictions. Our definition is consistent with the recent US Executive Order's notion of "foundation models with widely available model weights" (Executive Office of the President, 2023).

We consider this dichotomy because many of the risks described for open foundation models arise because developers relinquish exclusive control over downstream model use once model weights are released. For example, if developers impose restrictions on downstream usage, such restrictions are both challenging to enforce and easy for malicious actors to ignore. On the other hand, in the face of malicious use, developers of closed foundation models can, in theory, reduce, restrict, or block access to their models. In short, open release of model weights is irreversible.

As a result, some argue that widely available model weights could enable better research on their effects, promote competition and innovation, and improve scientific research, reproducibility, and transparency (Toma et al., 2023; Creative Commons et al., 2023; Cihon, 2023; Mozilla, 2023). Others argue that widely available model weights would enable malicious actors (Seger et al., 2023; Brundage et al., 2018) to more effectively misuse these models to generate disinformation (Solaiman et al., 2019b), non-consensual intimate imagery (Satter, 2023; Maiberg, 2023b), scams (Hazell, 2023), and bioweapons (Gopal et al., 2023; Soice et al.,

[2] Models are often accompanied by licenses that specify these terms. The Open Source Initiative designates some licenses, generally applied to code, as open source and is in the process of leading a similar effort to define open source AI. See: https://opensource.org/deepdive/

2023; Sandbrink, 2023; Matthews, 2023; Service, 2023; Bray et al., 2023). Appendix A provides a brief history of the debate on open foundation models.

3. Distinctive properties of open foundation models

Our work aims to better conceptualize the benefits and risks of open foundation models, especially in light of widespread disagreement within and beyond the AI community. Fundamentally, we decompose this into (i) identifying distinctive properties of open foundation models and (ii) reasoning about how those properties contribute to specific societal benefits and risks. Here, we enumerate five distinctive properties of open foundation models compared to closed foundation models. Note that other properties of foundation models, while not unique to open foundation models, may nonetheless influence the analysis of the benefits and risks of open foundation models. In particular, as models become more capable (Anderljung et al., 2023), these capabilities are likely to present new beneficial market opportunities but also greater misuse potential (e.g. more persuasive and targeted disinformation).

Broader access. Given our definition, open foundation models require that the model weights be widely available, if not to the public as a whole. While there may be some restrictions on who can use the model, given that such user restrictions are difficult to enforce or verify (as demonstrated by Meta's LLaMA 1 release in March 2023), model weights may effectively be available to the public. Functional barriers to use, ranging from requisite expertise to compute affordability, may nonetheless remain.

Greater customizability. By releasing model weights, open foundation models are readily customized for various downstream applications. Weights (and associated computational artifacts made available, such as activations and gradients) permit a wide range of adaptation methods for modifying the model, such as quantization (Frantar et al., 2023), fine tuning (Zhang et al., 2023; Dettmers et al., 2023), and pruning (Xia et al., 2023). While some closed foundation model developers permit certain adaptation methods (e.g. OpenAI allows fine tuning of GPT 3.5 as of January 2024), these methods tend to be more restrictive, costly, and ultimately constrained by the model developer's implementation. The customizability of open foundation models prevents model alignment interventions from being effective—such as by allowing users to fine-tune away alignment interventions (Narayanan et al., 2023), though similar issues also arise when closed models can be fine-tuned (Qi et al., 2023).

Local adaptation and inference ability. Users of an open foundation model can directly deploy it on local hardware, which removes the need for transferring data to the model developer. This allows for the direct use of the models without the need to share sensitive data with third parties, which is particularly important in sectors (e.g. finance, healthcare, law) where confidentiality and data protection are necessary—such as because of the sensitive nature of content or regulation around how data should be stored or transferred.

Inability to rescind model access. Once the weights for a foundation model are made widely available, little recourse exists for the foundation model developer to rescind access. While the foundation model developer, in coordination with distribution channels used to share model weights, can stop further access, existing copies of the model weights cannot be revoked. Furthermore, despite the developer's objections, users can redistribute model weights through, for example, peer-to-peer distribution (Vincent, 2023).

Inability to monitor or moderate model usage. For open foundation models, inference may be performed (i) locally (e.g. on a personal computer or self-owned cluster), (ii) on generic third-party computing platforms such as cloud services (e.g. Google Cloud Platform, Microsoft Azure), or (iii) on dedicated model hosting platforms (e.g. Together, Amazon Bedrock). In all cases, foundation model developers do not observe inference by default, making monitoring or moderation challenging, especially for local inference. Since dedicated model hosts are aware of what models are being used, developers may be able to coordinate with hosts to implement certain forms of monitoring/moderation.

4. Benefits of Open Foundation Models

Having established distinctive properties of open foundation models, we now critically analyze key benefits for open foundation models that emerge from these properties.

Distributing who defines acceptable model behavior. *Broader access and greater customizability expand who is able to specify the boundary of acceptable model behavior.*

Developers of closed foundation models exercise unilateral control in determining what is and is not acceptable model behavior. Given that foundation models increasingly intermediate critical societal processes (e.g. access to information, interpersonal communication; Lazar, 2023), much as social media platforms do today, the definition of what is acceptable model behavior is a consequential decision that should take into account the views of stakeholders and the context where the model is applied. In contrast, while developers may initially specify and control how the model responds to user queries, downstream developers who use open foundation models can modify them to specify alternative behavior. Open foundation models allow for greater diversity in defining what model behavior is accept-

able, whereas closed foundation models implicitly impose a monolithic view that is determined unilaterally by the foundation model developer.

Increasing innovation. *Broader access, greater customizability, and local inference expand how foundation models are used to develop applications.*

Since open foundation models can be more aggressively customized, they better support innovation across a range of applications. In particular, since adaptation and inference can be performed locally, application developers can more easily adapt or fine-tune models on large proprietary datasets without data protection and privacy concerns. Similarly, the customizability of open models allows improvements such as furthering the state-of-the-art across different languages (Pipatanakul et al., 2023). While some developers of closed foundation models provide mechanisms for users to opt out of data collection, the data storage, sharing, and usage practices of foundation model developers are not always transparent (Bommasani et al., 2023b).

However, the benefits of open foundation models for innovation may have limits due to potential comparative disadvantages in improving open foundation models over time. For example, open foundation model developers generally do not have access to user feedback and interaction logs that closed model developers do for improving models over time. Further, because open foundation models are generally more heavily customized, model usage becomes more fragmented and lessens the potential for strong economies of scale. However, new research directions such as merging models might allow open foundation model developers to reap some of these benefits (akin to open source software; Raffel, 2023). More generally, the usability of foundation models strongly influences innovation (Vipra & Korinek, 2023): factors beyond whether a model is released openly such as the capabilities of the model and the quality of potential inference APIs shape usability.

Accelerating science. *Broader access and greater customizability facilitate scientific research. The availability of other key assets (especially training data) would further accelerate scientific research.*

Foundation models are critical to modern scientific research, within and beyond the field of artificial intelligence. Broader access to foundation models enables greater inclusion in scientific research, and model weights are essential for several forms of research across AI interpretability, security, and safety (see Table A1). Ensuring ongoing access to specific models is essential for the scientific reproducibility of research, something that has been undermined to date by the business practice of closed model developers to retire models regularly (Kapoor & Narayanan, 2023). And since closed foundation models are often instrumented by safety measures by developers, these measures can complicate or render some research impossible. For example, Park et al. (2022) use foundation models without safety filters because their research aims to simulate human behavior (including toxic speech). Most closed foundation models, which are generally available through content-moderated APIs, would suppress these outputs.

However, model weights alone are insufficient for several forms of scientific research. Other assets, especially the data used to build the model, are necessary. For example, to understand how biases propagate, and are potentially amplified, requires comparisons of data biases to model biases, which in turn requires access to the training data (Wang & Russakovsky, 2021). Access to data and other assets, such as model checkpoints, has already enabled wide-ranging downstream research (Tian et al., 2023; Choi et al., 2023; Longpre et al., 2023b). While some projects prioritize accessibility to such assets with the stated goal of advancing scientific research on foundation models (Le Scao et al., 2022; Biderman et al., 2023), it is not common for open models in general. In fact, even the basic validity of model's evaluation depends on some transparency about the training data. For example, issues such as contamination might lead to overoptimistic results on benchmarks (Kapoor et al., 2024; Narayanan & Kapoor, 2023b). Access to information about the data can allow us to assess the amount of overlap between the training data and the test set.

Enabling transparency. *Broad access to weights enables some forms of transparency. The availability of other key assets (such as documentation and training data) would further improve transparency.*

Transparency is a vital precondition for responsible innovation and public accountability. Yet digital technologies are plagued by problematic opacity (see Bommasani et al., 2023b, §2.2). Widely available model weights enable external researchers, auditors, and journalists to investigate and scrutinize foundation models more deeply. In particular, such inclusion is especially valuable given that the foundation model developers often underrepresent marginalized communities that are likely to be subject to the harms of foundation models. The history of digital technology demonstrates that broader scrutiny, including by those belonging to marginalized groups that experience harm most acutely, reveals concerns missed by developers (Sweeney, 2013; Noble, 2018; Buolamwini & Gebru, 2018; Raji & Buolamwini, 2019). The Foundation Model Transparency Index as of May 2024 indicates that developers of major open foundation models tend to be more transparent than their closed counterparts (Bommasani et al., 2024).

Still, model weights only make some types of transparency (e.g. evaluations of risk) possible, but they do not guarantee such transparency will manifest. More generally,

model weights do not guarantee transparency on the upstream resources used to build the foundation model (e.g. data sources, labor practices, energy expenditure) nor transparency on the downstream impact of the foundation model (e.g. affected markets, adverse events, usage policy enforcement). Such transparency can help address prominent societal concerns surrounding bias (Birhane et al., 2023; Thiel, 2023), privacy (Ippolito et al., 2023), copyright (Henderson et al., 2023; Lee et al., 2023; Longpre et al., 2023a), labor (Perrigo, 2023; Hao & Seetharaman, 2023), usage practices (Narayanan & Kapoor, 2023a), and demonstrated harms (Guha et al., 2023).

Mitigating monoculture and market concentration. *Greater customizability mitigates the harms of monoculture and broader access reduces market concentration.*

Foundation models function as infrastructure for building downstream applications, spanning market sectors (Bommasani et al., 2021; 2023c; Vipra & Korinek, 2023; UK CMA, 2023). By design, they contribute to the rise of algorithmic monoculture (Kleinberg & Raghavan, 2021; Bommasani et al., 2022): many downstream applications depend on the same foundation model. Monocultures often yield poor societal resilience and are susceptible to widespread systemic risk: consider the Meltdown and Spectre attacks, which led to massive security risks because of the widespread dependence on Intel and ARM-based microprocessors (Kocher et al., 2018; Lipp et al., 2018; Staff, 2018). Further, foundation model monocultures have been conjectured to lead to correlated failures (Bommasani et al., 2022) and cultural homogenization (Lee et al., 2022; Padmakumar & He, 2023). Since open foundation models are more easily customized, they may yield more diverse downstream model behavior, thereby reducing the severity of homogeneous outcomes.

Broad access to model weights and greater customizability further enable greater competition in downstream markets, helping to reduce market concentration at the foundation model level from vertical cascading. In the foundation model market, there are barriers to entry for low-resource actors in developing foundation models given their significant capital costs (Vipra & Korinek, 2023; UK CMA, 2023). For example, training the Llama 2 series of models required 3.3 million GPU hours on NVIDIA A100-80GB GPUs (Touvron et al., 2023b): at February 2024 cloud computing rates of $1.8/GPU hour (Lambda, 2024), training this model would cost around $6 million. Further, while open foundation models may increase competition in some regions of the AI supply chain, they are unlikely to reduce market concentration in the highly concentrated upstream markets of computing and specialized hardware providers (Widder et al., 2023).

5. Risks of Open Foundation Models

Technologists and policymakers have worried that open foundation models present risks, in particular, due to the inability to monitor, moderate, or revoke access. We survey the literature on misuse vectors specifically associated with open foundation models, identifying biosecurity, cybersecurity, voice cloning scams, spear phishing, disinformation, non-consensual intimate imagery, and child sexual abuse material (Seger et al., 2023; Thiel et al., 2023; Maiberg, 2023a).[3] To understand the nature of these risks, we present a framework that centers *marginal* risk: what additional risk is society subject to because of open foundation models relative to pre-existing technologies, closed models, or other relevant reference points?

5.1. Risk Assessment Framework

To assess the risk of open foundation models for a specific misuse vector, we present a six-point framework. Underpinning this is an emphasis on communicating assumptions and uncertainty: misuse vectors often involve complex supply chains and the capabilities of foundation models are rapidly evolving, meaning the balance of power between attackers and defenders can be unstable (Shevlane & Dafoe, 2020).

The risk framework enables precision in discussing the misuse risk of open foundation models and is based on the threat modeling framework in computer security (Drake, 2021; Shostack, 2014; Crothers et al., 2023; Seaman, 2022; Drake, 2021). For example, without clearly articulating the marginal risk of biosecurity concerns stemming from the use of open (natural) language models, researchers might come to completely different conclusions about whether they pose risks: open language models can generate accurate information about pandemic-causing pathogens (Gopal et al., 2023), yet such information is publicly available on the Internet, even without the use of open language models (Guha et al., 2023).[4]

1. Threat identification. All misuse analyses should systematically identify and characterize the potential threats being analyzed (Shostack, 2014; Crothers et al., 2023; Seaman, 2022; Drake, 2021). In the context of open foundation models, this would involve naming the misuse vector, such as spear-phishing scams or influence operations, as well as

[3] Some have also discussed that (open) foundation models may contribute to existential risk via speculative AI takeover scenarios, which we do not consider here.

[4] In addition, two recent studies found that access to language models does not significantly increase access to information required to carry out biosecurity attacks compared to Internet access (Mouton et al., 2024; Patwardhan et al., 2024). More importantly, access to information might not be a major barrier for carrying out such attacks—stronger interventions might lie downstream (Batalis, 2023).

Misuse risk	Paper	Threat identification	Existing risk (absent open FMs)	Existing defenses (absent open FMs)	Evidence of marginal risk	Ease of defense	Uncertainty/assumptions
Spear-phishing scams	Hazell (2023)	●	◑	○	○	◑	○
Cybersecurity risk	Seger et al. (2023)	◑	○	◑	○	◑	○
Disinformation	Musser (2023)	●	◑	○	○	◑	●
Biosecurity risk	Gopal et al. (2023)	●	○	◑	○	○	○
Voice-cloning scams	Ovadya et al. (2019)	●	◑	◑	◑	◑	●
Non-consensual intimate imagery	Lakatos (2023)	●	◑	○	◑	◑	○
Child sexual abuse material	Thiel et al. (2023)	●	●	●	●	●	●

Table 1. Misuse analyses of open foundation models assessed under our risk framework (§5.1). ● indicates the step of our framework is clearly addressed; ◑ indicates partial completion; ○ indicates the step is absent in the misuse analysis. Incomplete assessments do not indicate that the analysis in prior studies is flawed, only that these studies, on their own, do not show an increased marginal societal risk stemming from open foundation models. We provide more details for our assessment of each row in Appendix B.

detailing the manner in which the misuse would be executed. To present clear assumptions, this step should clarify the potential malicious actors and their resources: individual hackers are likely to employ different methods and wield different resources relative to state-sponsored entities.

2. Existing risk (absent open foundation models). Given a threat, misuse analyses should clarify the existing misuse risk in society. For example, Seger et al. (2023) outline the misuse potential for open foundation models via disinformation on social media, spear-phishing scams over email, and cyberattacks on critical infrastructure. Each of these misuse vectors already are subject to risk *absent* open foundation models: understanding the pre-existing level of risk contextualizes and baselines any new risk introduced by open foundation models.

3. Existing defenses (absent open foundation models). Assuming that risks exist for the misuse vector in question, misuse analyses should clarify how society (or specific entities or jurisdictions) defends against these risks. Defenses can include technical interventions (e.g. spam filters to detect and remove spear-phishing emails) and regulatory interventions (e.g. laws punishing the distribution of child sexual abuse material). Understanding the current defensive landscape informs the efficacy, and sufficiency, with which new risks introduced by open foundation models will be addressed.

4. Evidence of marginal risk of open FMs. The threat identification, paired with an analysis of existing risks and defenses, provides the conceptual foundation for reasoning about the risks of open foundation models. Namely, subject to the status quo, we can evaluate the *marginal risk* of open foundation models. Being aware of existing risk clarifies instances where open foundation models simply duplicate existing risk (e.g. an open language model providing biological information available via Wikipedia). Similarly, being aware of existing defenses clarifies instances where open foundation models introduce concerns that are well-addressed by existing measures (e.g., email and OS-based filters detecting spear-phishing emails, whether human or AI-generated; CraigMarcho, 2007; Apple Support, 2023; Google, 2023). Conversely, we can identify critical instances where new risks are introduced (e.g. fine tuning models to create non-consensual intimate imagery of specific people; see Table 2; Maiberg, 2023b) or where existing defenses will be inadequate (e.g. AI-generated child sexual abuse material may overwhelm existing law enforcement resources; Harwell, 2023).

Further, the marginal risk analysis need not only be conducted relative to the status quo, but potentially relative to other (possibly hypothetical) baselines. For example, understanding the marginal risk of open release relative to a more restricted release (e.g. API release of a closed foundation model) requires reasoning about the relevant existing defenses for said restricted release. This perspective ensures greater care is taken to not assume that closed releases are intrinsically more safe and, instead, to interrogate the quality of existing defenses (e.g. the fallibility of existing API safeguards; Qi et al., 2023).

5. Ease of defending against new risks. While existing defenses provide a baseline for addressing new risks introduced by open foundation models, they do not fully clarify the marginal risk. In particular, new defenses can be implemented or existing defenses can be modified to address the increase in overall risk. Therefore, characterizations of the marginal risk should anticipate how defenses will evolve in reaction to risk: for example, (open) foundation models may also contribute to such defenses (e.g. the creation of better disinformation detectors; Zellers et al. (2019) or code fuzzers; Liu et al. (2023)).

6. Uncertainty and assumptions. Finally, it is imperative

Framework step	**Cybersecurity** Automated vulnerability detection	**Non-consensual intimate imagery (NCII)** Digitally altered NCII
Threat identification	Vulnerability detection tools can be used to automate the process of discovering software vulnerabilities. Threat actors include individual hackers, small groups, or state-sponsored attackers.	Digital tools can be used to alter images of people without their consent in sexually explicit ways. Threat actors are typically individuals or coordinated groups (such as on online platforms like Reddit or Telegram) creating imagery of people they know as well as public figures.
Existing risk (absent open FMs)	Attackers benefit from the natural worst-case asymmetry in vulnerability detection: attackers need to exploit only a single effective vulnerability to succeed, whereas defenders must defend against all vulnerabilities to succeed. Existing risk is heavily influenced by the resources of the attacker: sophisticated attackers often make use of automated vulnerability detection tools in attack design. Fuzzing tools have long been used to find vulnerabilities in software (Takanen et al., 2008), as have tools like Metasploit, a free penetration testing framework that can aid automated vulnerability detection (Kennedy et al., 2011). MITRE's Adversarial Threat Landscape for Artificial-Intelligence Systems, a cybersecurity threat matrix for adversarial machine learning, includes many techniques that make use of closed foundation models and other types of machine learning models to detect vulnerabilities (MITRE, 2021).	Photoshop has long been used to create digitally altered NCII (Broughton, 2009). In the last decade, tools to create NCII using face swapping and other rudimentary ML techniques have become popular (Widder et al., 2022). A telegram bot that used such techniques was used to generate over 100,000 sexualized images of women (Ajder et al., 2020). Digitally altered NCII and also be used to extort victims (Joshi, 2021; Satter, 2023), in addition to its emotional and psychological tolls (Roberts, 2019; Scott, 2020; Hao, 2021).
Existing defenses (absent open FMs)	Cybersecurity defenses often adopt defense-in-depth strategies, where defenses are layered to ensure an exploit based on an unaddressed vulnerability in one layer does not affect other layers of defenses (Kuipers & Fabro, 2006). Within the vulnerability detection setting, defenders can preemptively use vulnerability detection tools to detect and patch security threats, again dependent on their access to resources. Incentive strategies, such as bug bounties, can tilt the offense-defense balance in favor of defense to some extent by incentivizing bug finders (hackers, security researchers, firms) to report vulnerabilities.	The software for creating digitally altered NCII can run on consumer-grade devices and has proliferated widely. There are efforts to reduce the use of such tools for creating NCII in open source communities (Widder et al., 2022), but these efforts are unlikely to be sufficient since there are several mechanisms for accessing the software. However, online platforms where NCII is distributed, such as social media platforms, can take steps to curb its spread (Thiel & Einstein, 2020). For example, a nonprofit called Stop NCII coordinates takedowns of known NCII across online platforms (Mortimer, 2021).
Evidence of marginal risk of open FMs	We are unaware of existing evidence that malicious users have successfully used open foundation models to automate vulnerability detection. Dark web advertisements for tools exist, claiming to facilitate automated vulnerability detection, but it is unclear if these products rely on open FMs (Amos, 2023). In considering marginal risks relative to closed foundations, while closed foundation models can be better monitored for misuse, it is not clear if such uses will be reliability identified. Namely, using a closed foundation model for vulnerability detection is not necessarily misuse, which introduces a nontrivial classification problem of distinguishing between legitimate and malicious uses of closed foundation models for automated vulnerability detection (see Figures A1a and A1b).	Over the last two years, open FMs have been used for creating vast amounts of digitally altered NCII. Compared to previous tools for creating sexualized imagery, open FMs can be fine tuned to create sexualized images of specific people (Maiberg, 2023b). Compared to using tools like Photoshop, once such a fine-tuned model is made available, it is much easier for nonexperts to use these tools. While developers of closed FMs can enforce guardrails on the use of their text-to-image models for creating NCII, such guardrails on open FMs can be easily circumvented. There have been several real-world incidents involving the use of open FMs for creating NCII, leading to clear, demonstrated harm (Llach, 2023; Cañas, 2023; Kaspersky, 2023).
Ease of defense	Similar to previous waves of automated vulnerability detection, LLMs can be incorporated into the information security toolkit to bolster defense. For example, Liu et al. (2023) show how LLMs can expand the coverage of a popular fuzzing tool called OSS-Fuzz. Foundation models can be used to monitor signals from deployed software systems for signs of active exploits as well. Google has made use of LLMs in its popular malware detection platform VirusTotal, using models to help explain the functionality of malware contained in a particular file (Quintero, 2023). Defense in depth will continue to remain important in aiding defense. Regardless of whether the model used for automated vulnerability detection is open or proprietary, signals and the ability to analyze them at machine scale and speed differentially supports defenders because of better access to the systems.	Open FMs used to create NCII require few resources to run—indeed, many prominent text-to-image models can run on an iPhone or MacBook. As a result, non-proliferation of these models is generally not feasible. In contrast, crackdowns on the distribution of specifically tailored models for creating NCII is feasible and warranted, as is distribution of the content (Gorwa & Veale, 2023; Maiberg, 2023a). There are several legislative proposals to penalize the creation and distribution of digitally altered NCII, though given that channels for the spread of NCII can be anonymous or end-to-end encrypted, the efficacy of such legislation remains to be seen (Illinois General Assembly, 2023; Saliba, 2023; Reid, 2020; Kocsis, 2021; Hao, 2021; Siddique, 2023).
Uncertainty and assumptions	The analysis of marginal risk and ease of defense assumes that defenders will continue to have better access to state-of-the-art vulnerability detection tools, including those based on open FMs. It also assumes investment by defenders in using these tools to update their infosec practices and that the offense-defense balance will not change dramatically as the capabilities of models improve.	Technical solutions for curtailing the use of already existing models to create NCII are hard or impossible. Even if future models can have robust technical safeguards, already-released models will continue to be misused. Even if downstream providers take steps to moderate digitally altered NCII, misuse can happen on smaller (anonymous/end-to-end encrypted) platforms or on the dark web by malicious users.

Table 2. Instantiation of our risk analysis framework for cybersecurity (automated vulnerability detection) and non-consensual intimate imagery (digitally altered NCII).

to articulate the uncertainties and assumptions that underpin the risk assessment framework for any given misuse risk. This may encompass assumptions related to the trajectory of technological development, the agility of threat actors in adapting to new technologies, and the potential effectiveness of novel defense strategies. For example, forecasts of how model capabilities will improve or how the costs of model inference will decrease would influence assessments of misuse efficacy and scalability.

Using our risk assessment framework, we assess past studies that span different risk vectors in Table 1. We find that the risk analysis is incomplete for six of the seven studies we analyze. To be clear, incomplete assessments do not necessarily indicate that the analysis in prior studies is flawed, only that these studies, on their own, are insufficient evidence to demonstrate increased marginal societal risk from open foundation models.

In Table 2, we instantiate the framework for two misuse risks, providing preliminary analyses of cybersecurity risks stemming from automated vulnerability detection and the risk of digitally altered NCII. For the former, we find that the current marginal risk of open foundation models is low and that there are several approaches to defending against the marginal risk, including using AI for defense. For the latter, open foundation models pose considerable marginal risk at present, and plausible defenses seem hard. Note that these are not the only risks from foundation models (Barrett et al., 2023)—for example, the creation of malware is another cybersecurity risk that requires separate analysis—yet when researchers talk about cybersecurity risks of open foundation models, they often club together different threats. This illustrates how the framework helps clarify the points of contention in debates on open foundation models. Critically, while many of the same properties of open foundation models are relevant for analyzing different misuse vectors (such as the inability to revoke access), the risk assessment framework helps introduce specifics that differentiate the misuse vector, for instance, by pointing out elements of the misuse supply chain where risk is better addressed.

As the capabilities of foundation models (including open models) improve, the risk assessment framework can guide analyses of societal risks from increasing capability by providing a grounded analysis of whether model releases bring about increased marginal risk to society. Still, it is important to note the limitations on the scope of the framework's applicability. First, while the risk assessment framework can help clarify the societal risks of releasing a foundation model openly, note that it is not a complete framework for making release decisions since it does not provide a mechanism for trading the marginal benefits of openly releasing models against the marginal risk, nor does it look at the opportunity cost of *not* releasing a model openly. Second, while the framework allows an evaluation of the risk of releasing models openly for known risks (such as cybersecurity, biosecurity etc.), it does not account for *unknown unknowns*—risks that we have no prior understanding of. Third, there could be a number of coordination issues among actors for figuring out when to release models—for example, to reduce the risk of NCII, open model developers would need to coordinate with social media platforms as well as other downstream platforms like CivitAI (see Table 2). While the framework allows us to identify such opportunities, it does not automatically bring about the coordination of these actors. Overall, while the framework improves the precision, rigor, and completeness of risk assessment, we expect other approaches to analyzing risk will be needed for addressing these limitations.

6. Recommendations and calls to action

Armed with a clearer conceptualization of the benefits and a framework for assessing the risks of open foundation models, we make the following recommendations to (i) AI developers, (ii) researchers investigating AI risks, (iii) policymakers, and (iv) competition regulators.

AI developers. In contrast to closed foundation models, which are usually treated by their developers and their users with product safety expectations, open foundation models have less clear safety expectations. In particular, the division of responsibility for safety between the developer and user of an open foundation model is unclear and lacks established norms. Consequently, developers of open foundation models should be transparent about both the responsible AI practices they implement and the responsible AI practices they recommend or delegate to downstream developers or deployers. In turn, when downstream users are procuring foundation models, they should consider which responsible AI measures have already been implemented (and their efficacy if measured) and, accordingly, implement or demand responsible AI practices from developers (Vinsel, 2019). This would help ensure that responsible AI practices do not fall through the cracks as providers of downstream AI applications leverage open foundation models from other upstream providers.

Researchers investigating AI risks. Our preliminary analysis of the misuse risk of open foundation models reveals significant uncertainty for several misuse vectors due to incomplete or unsatisfactory evidence. In turn, researchers investigating AI risks should conduct new research to clarify the marginal risks for misuse of open foundation models. In particular, in light of our observations regarding past work, greater attention should be placed on articulating the status quo, constructing realistic threat models (or arguments for why speculative threat models yield generalizable evidence), and considering the full supply chain for misuse.

Policymakers. Government funding agencies should ensure that research investigating the risks of open foundation models is sufficiently funded while remaining appropriately independent from the interests of foundation model developers (Lucas et al., 2023). Once the uncertainty around specific misuse vectors is reduced (including via improved tracing of downstream model usage), and if the marginal risks are shown to be significant enough to warrant concern, further policy interventions (e.g. hardening downstream attack surfaces) can be considered. Policymakers should also proactively assess the impacts of proposed regulation on developers of open foundation models (Bankston & Hodges, 2024). In particular, some policy proposals impose high compliance burdens for these developers, and such policies should only be pursued with sufficient justification of the adverse effect on the open foundation model ecosystem. Policies that place obligations on foundation model developers to be responsible for downstream use are intrinsically challenging, if not impossible, for open developers to meet. If recent proposals for liability (Blumenthal & Hawley, 2023b) and watermarking (Executive Office of the President, 2023; Chinese National Information Security Standardization Technical Committee, 2023; G7 Hiroshima Summit, 2023) are interpreted strictly to apply to foundation model developers, independent of how the model is adapted or used downstream, they would be difficult for open developers to comply with (Bommasani et al., 2023a), since these developers have little ability to monitor, moderate, or prohibit downstream usage.

Competition regulators. Significant theoretical benefits of open foundation models relate to their potential to catalyze innovation, distribute power, and foster competition. With this in mind, the magnitude of these economic benefits is largely undocumented in the absence of large-scale economic analyses or market surveillance. For example, many benefits hinge on open foundation models meaningfully expanding consumer choice and reducing costs. If factors such as differences in model quality are the more direct causes of the adoption of specific foundation models, these purported benefits may not manifest. In turn, competition regulators should invest in measuring the benefits of foundation models and the impact of openness on those benefits. In particular, the UK's Competition and Markets Authority has begun such work (UK CMA, 2023), which would be bolstered by parallel efforts across other jurisdictions.

7. Conclusion

Open foundation models are controversial due to fundamental philosophical disagreements, fragmented conceptual understanding, and poor empirical evidence. Our work aims to rectify the conceptual confusion by clearly defining open foundation models, identifying their distinctive properties, and clarifying their benefits and risks. While it is unlikely that certain underlying philosophical tensions will ever be resolved, especially when inextricably intertwined with the incentives of different actors in the AI space, we encourage future work to address today's deficits in empirical evidence. Overall, we are optimistic that open foundation models can contribute to a vibrant AI ecosystem, but realizing this vision will require significant action from many stakeholders.

Impact Statement

Our work explicitly addresses the societal impact of open foundation models. In turn, we expect that our work is likely to influence different stakeholders to take the actions we call for in §6. For example, by articulating specific benefits and risks and, more broadly, providing the conceptual framework for reasoning about these matters, we anticipate that policymakers may take action per our characterization of open foundation models.

Beyond these impacts, we specifically raise the concern of misinterpretation of our work. Throughout, as we emphasize in our risk assessment framework, we highlight areas where we are uncertain (e.g. there may be little evidence we are aware of for certain threat vectors or current work does not suffice for showing marginal risks of open foundation models). Further, as we note, foundation models and the capabilities/properties of foundation models that are openly released will change over time. For all of these reasons, we caution against over-indexing on specific characterizations of the current benefit vs. risk balance.

Finally, we bring attention to the concern that this work further entrenches a salient divide in the community of "open vs. closed". While we do see substantive reasons for distinguishing widely available model weights for the purposes of analysis of societal impact, this distinction is not the only useful one (e.g. improved transparency and scientific quality, understanding of biases and copyright may hinge on open data). Regardless, we believe there is significant room for consensus on the current status quo, or the conceptual framework, as opposed to the current fierce debate. And, most critically, many aspects of the societal impact of foundation models are largely orthogonal to the open vs. closed distinction and should not be overlooked as a result. For example, many matters pertaining to how a foundation model is built (e.g. consent for the use of data, copyright and intellectual property considerations, presence of sensitive information, energy consumption and environmental impacts), and how foundation models are used in practice once made available like labor displacement, may apply similarly to both open and closed foundation models.

Acknowledgements

We thank Ellie Evans, Helen Toner, Ion Stoica, Marietje Schaake, Nate Persily, Nicholas Carlini, Rob Reich, Steven Cao, and Zico Kolter for extensive feedback on this work. We are grateful to the participants of the Stanford Workshop on the Governance of Open Foundation Models and the Princeton-Stanford Workshop on Responsible and Open Foundation Models for their feedback and input. This work was supported in part by the AI2050 program at Schmidt Futures (Grant G-22-63429), the Patrick J. McGovern Foundation, and the Hoffman-Yee grant program of the Stanford Institute for Human-Centered Artificial Intelligence.

References

Ajder, H., Patrini, G., and Cavalli, F. Automating Image Abuse: Deepfake Bots on Telegram. *Sensity*, October 2020.

Amos, Z. What Is FraudGPT?, August 2023. URL https://hackernoon.com/what-is-fraudgpt.

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., Schuett, J., Shavit, Y., Siddarth, D., Trager, R., and Wolf, K. Frontier AI Regulation: Managing Emerging Risks to Public Safety, November 2023. URL http://arxiv.org/abs/2307.03718. arXiv:2307.03718 [cs].

Apple Support. Safely open apps on your Mac, September 2023. URL https://support.apple.com/en-us/HT202491.

Bankston, K. and Hodges, J. Openness and transparency in ai provide significant benefits for society, 2024. URL https://cdt.org/wp-content/uploads/2024/03/Civil-Society-Letter-on-Openness-for-NTIA-Process-March-25-2024.pdf.

Barrett, C., Boyd, B., Bursztein, E., Carlini, N., Chen, B., Choi, J., Chowdhury, A. R., Christodorescu, M., Datta, A., Feizi, S., Fisher, K., Hashimoto, T., Hendrycks, D., Jha, S., Kang, D., Kerschbaum, F., Mitchell, E., Mitchell, J., Ramzan, Z., Shams, K., Song, D., Taly, A., and Yang, D. Identifying and Mitigating the Security Risks of Generative AI. *Foundations and Trends® in Privacy and Security*, 6(1):1–52, December 2023. ISSN 2474-1558, 2474-1566. doi: 10.1561/3300000041. URL https://www.nowpublishers.com/article/Details/SEC-041. Publisher: Now Publishers, Inc.

Basdevant, A. et al. Towards a framework for openness in foundation models. Proceedings from the Columbia Convening on Openness in Artificial Intelligence, May 2024.

Batalis, S. Can Chatbots Help You Build a Bioweapon?, November 2023. URL https://foreignpolicy.com/2023/11/05/ai-artificial-intelligence-chatbot-bioweapon-virus-bacteria-genetic-engineering/.

Biderman, S. Good Work Enabled by Open Models, November 2023. URL https://docs.google.com/spreadsheets/d/1kt5jp1U50AfDGEKsqFXvtCCQ8HYvQGcya1vf_OaD8x0/edit?usp=sharing&usp=embed_facebook.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Birhane, A., Prabhu, V., Han, S., Boddeti, V. N., and Luccioni, A. S. Into the LAIONs Den: Investigating Hate in Multimodal Datasets, November 2023. URL http://arxiv.org/abs/2311.03449. arXiv:2311.03449 [cs].

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., et al. Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136, 2022.

Blumenthal, R. and Hawley, J. Letter on the leak of meta's ai model. *US Senate*, 2023a. URL https://www.blumenthal.senate.gov/imo/media/doc/06062023metallamamodelleakletter.pdf.

Blumenthal, R. and Hawley, J. Bipartisan Framework for U.S. AI Act, September 2023b. URL https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto,

T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., and Liang, P. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems*, 2022.

Bommasani, R., Kapoor, S., Klyman, K., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., and Liang, P. Considerations for Governing Open Foundation Models, December 2023a. URL https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., and Liang, P. The foundation model transparency index, 2023b.

Bommasani, R., Soylu, D., Liao, T., Creel, K. A., and Liang, P. Ecosystem graphs: The social footprint of foundation models. *ArXiv*, abs/2303.15772, 2023c. URL https://api.semanticscholar.org/CorpusID:257771875.

Bommasani, R., Klyman, K., Kapoor, S., Longpre, S., Xiong, B., Maslej, N., and Liang, P. The foundation model transparency index v1.1. May 2024. URL https://arxiv.org/abs/2310.12941.

Bray, H., June 28, U., 2023, and Comments5, . a. m. S. o. F. S. o. T. AI might help unleash the next pandemic, MIT study says - The Boston Globe, 2023. URL https://www.bostonglobe.com/2023/06/28/business/ai-might-help-unleash-next-pandemic-mit-study-says/.

Broughton, A. Tennessee man charged in 'virtual pornography' case, June 2009. URL https://edition.cnn.com/2009/CRIME/06/24/virtual.child.porn/index.html.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Sean O hEigeartaigh, Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., and Amodei, D. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, February 2018. URL http://arxiv.org/abs/1802.07228. arXiv:1802.07228 [cs].

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Cañas, J. Illegal use of artificial intelligence to create deepfakes depicting sexual content of minors. *European Parliament*, 2023. URL https://www.europarl.europa.eu/doceo/document/E-9-2023-002788_EN.html.

Chesney, R. and Citron, D. K. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, July 2018. URL https://papers.ssrn.com/abstract=3213954.

Chinese National Information Security Standardization Technical Committee. Basic Safety Requirements for Generative Artificial Intelligence Services (Draft for Feedback), November 2023. URL https://cset.georgetown.edu/publication/china-safety-requirements-for-generative-ai/.

Choi, D., Shavit, Y., and Duvenaud, D. Tools for verifying neural models' training data. *arXiv preprint arXiv:2307.00682*, 2023.

Cihon, P. How to get AI regulation right for open source, July 2023. URL https://github.blog/2023-07-26-how-to-get-ai-regulation-right-for-open-source/.

Cole, S. AI-Assisted Fake Porn Is Here and We're All Fucked, December 2017. URL https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn.

CraigMarcho. IE7 - Introducing the Phishing Filter, March 2007. URL https://techcommunity.microsoft.com/t5/ask-the-performance-team/ie7-introducing-the-phishing-filter/ba-p/372327.

Creative Commons, EleutherAI, GitHub, Hugging Face, LAION, and Open Future. Supporting Open Source and Open Science in the EU AI Act. Technical report, July 2023. URL https://github.blog/wp-content/uploads/2023/07/Supporting-Open-Source-and-Open-Science-in-the-EU-AI-Act.pdf.

Crothers, E. N., Japkowicz, N., and Viktor, H. L. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002, 2023. doi: 10.1109/ACCESS.2023.3294090.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

Drake, V. Threat Modeling, August 2021. URL https://owasp.org/www-community/Threat_Modeling.

Executive Office of the President. Safe, secure, and trustworthy development and use of artificial intelligence. Executive Order, 10 2023. URL https://www.federalregister.gov/documents/2023/10/30/2023-24110/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence. Federal Register Vol. 88, No. 210 (October 30, 2023).

Flavio Ceolin. Two buffer overflow vulnerabilities in Zephyr USB code, 2023. URL https://github.com/zephyrproject-rtos/zephyr/security/advisories/GHSA-4vgv-5r6q-r6xh.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. OPTQ: Accurate Quantization for Generative Pre-trained Transformers. 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

G7 Hiroshima Summit. Hiroshima Process International Guiding Principles for Advanced AI system. Technical report, October 2023. URL https://www.mofa.go.jp/files/100573471.pdf.

Google. Email sender guidelines, December 2023. URL https://support.google.com/mail/answer/81126?hl=en.

Gopal, A., Helm-Burger, N., Justen, L., Soice, E. H., Tzeng, T., Jeyapragasan, G., Grimm, S., Mueller, B., and Esvelt, K. M. Will releasing the weights of future large language models grant widespread access to pandemic agents?, November 2023. URL http://arxiv.org/abs/2310.18233. arXiv:2310.18233 [cs].

Gorwa, R. and Veale, M. Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries, November 2023. URL http://arxiv.org/abs/2311.12573. arXiv:2311.12573 [cs].

Guha, N., Lawrence, C. M., Gailmard, L. A., Rodolfa, K. T., Surani, F., Bommasani, R., Raji, I. D., Cuéllar, M.-F., Honigsberg, C., Liang, P., and Ho, D. E. Ai regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *George Washington Law Review, Symposium on Legally Disruptive Emerging Technologies*, 2023.

Han, T., Adams, L. C., Papaioannou, J.-M., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., and Bressem, K. K. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

Hao, K. Deepfake porn is ruining women's lives. Now the law may finally ban it., February 2021. URL https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/.

Hao, K. and Seetharaman, D. Cleaning up chatgpt takes heavy toll on human workers. *The Wall Street Journal*, July 2023. URL https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483. Photographs by Natalia Jidovanu.

Harwell, D. AI-generated child sex images spawn new nightmare for the web. *Washington Post*, June 2023. ISSN 0190-8286. URL https://www.washingtonpost.com/technology/2023/06/19/artificial-intelligence-child-sex-abuse-images/.

Hazell, J. Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns, May 2023. URL http://arxiv.org/abs/2305.06972. arXiv:2305.06972 [cs].

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use. *arXiv preprint arXiv:2303.15715*, 2023.

Illinois General Assembly. Full Text of HB2123, July 2023. URL https://www.ilga.gov/legislation/fulltext.asp?DocName=&SessionId=112&GA=103&DocTypeId=HB&DocNum=2123&GAID=17&LegID=145586&SpecSess=&Session=.

Ippolito, D., Tramer, F., Nasr, M., Zhang, C., Jagielski, M., Lee, K., Choquette Choo, C., and Carlini, N. Preventing Generation of Verbatim Memorization in Language Models Gives a False Sense of Privacy. In Keet, C. M., Lee, H.-Y., and Zarrieß, S. (eds.), *Proceedings of the 16th International Natural Language Generation Conference*, pp. 28–53, Prague, Czechia, September 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.inlg-main.3. URL https://aclanthology.org/2023.inlg-main.3.

Joshi, S. They Follow You on Instagram, Then Use Your Face To Make Deepfake Porn in This Sex Extortion Scam, September 2021. URL https://www.vice.com/en/article/z3x9yj/india-instagram-sextortion-phishing-deepfake-porn-scam.

Kapoor, S. and Narayanan, A. OpenAI's policies hinder reproducible research on language models, March 2023. URL https://www.aisnakeoil.com/p/openais-policies-hinder-reproducible.

Kapoor, S., Henderson, P., and Narayanan, A. Promises and pitfalls of artificial intelligence for legal applications, January 2024. URL http://arxiv.org/abs/2402.01656. arXiv:2402.01656 [cs].

Kaspersky. How real is deepfake threat? *Kaspersky Daily*, 2023.

Kennedy, D., O'Gorman, J., Kearns, D., and Aharoni, M. *Metasploit: The Penetration Tester's Guide*. No Starch Press, USA, 1st edition, 2011. ISBN 159327288X.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A Watermark for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 17061–17084. PMLR, July 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html. ISSN: 2640-3498.

Kleinberg, J. and Raghavan, M. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021. doi: 10.1073/pnas.2018340118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2018340118.

Kocher, P., Genkin, D., Gruss, D., Haas, W., Hamburg, M., Lipp, M., Mangard, S., Prescher, T., Schwarz, M., and Yarom, Y. Spectre Attacks: Exploiting Speculative Execution, January 2018. URL http://arxiv.org/abs/1801.01203. arXiv:1801.01203 [cs].

Kocsis, E. Deepfakes, Shallowfakes, and the Need for a Private Right of Action Comments. *PennState Law Review*, 126(2):621–650, 2021. URL https://heinonline.org/HOL/P?h=hein.journals/dknslr126&i=633.

Kuipers, D. and Fabro, M. Control Systems Cyber Security:Defense in Depth Strategies. Technical Report INL/EXT-06-11478, Idaho National Lab. (INL), Idaho Falls, ID (United States), May 2006. URL https://www.osti.gov/biblio/911553.

Lakatos, S. A Revealing Picture: AI-Generated 'Undressing' Images Move from Niche Pornography Discussion Forums to a Scaled and Monetized Online Business. Technical report, December 2023. URL https://public-assets.graphika.com/reports/graphika-report-a-revealing-picture.pdf.

Lambda. GPU Cloud - VMs for Deep Learning, February 2024. URL https://web.archive.org//web/20240226155153/https://lambdalabs.com/service/gpu-cloud#pricing.

Lazar, S. Governing the algorithmic city. *Tanner Lectures*, 2023. URL https://write.as/sethlazar/.

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P. O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A. F., Alfassy, A., Rogers, A., Nitzav, A. K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., Adelani, D. I., Radev, D., Ponferrada, E. G., Levkovizh, E., Kim, E., Natan, E. B., De Toni, F., Dupont, G., Kruszewski, G., Pistilli, G., Elsahar, H., Benyamina, H., Tran, H., Yu, I., Abdulmumin, I., Johnson, I., Gonzalez-Dios, I., de la Rosa, J., Chim, J., Dodge, J., Zhu, J., Chang, J., Frohberg, J., Tobing, J., Bhattacharjee, J., Almubarak, K., Chen, K., Lo, K., Von Werra, L., Weber, L., Phan, L., allal, L. B., Tanguy, L., Dey, M., Muñoz, M. R., Masoud, M., Grandury, M., Šaško, M., Huang, M., Coavoux, M., Singh, M., Jiang, M. T.-J., Vu, M. C., Jauhar, M. A., Ghaleb, M., Subramani, N., Kassner, N., Khamis, N., Nguyen, O., Espejel, O., de Gibert, O., Villegas, P., Henderson, P., Colombo, P., Amuok, P., Lhoest, Q., Harliman, R., Bommasani, R., López, R. L., Ribeiro, R., Osei, S., Pyysalo, S., Nagel, S., Bose, S., Muhammad, S. H., Sharma, S., Longpre, S., Nikpoor, S., Silberberg, S., Pai, S., Zink, S., Torrent, T. T., Schick, T., Thrush, T., Danchev, V., Nikoulina, V., Laippala, V., Lepercq, V., Prabhu, V., Alyafeai, Z., Talat, Z., Raja, A., Heinzerling, B., Si, C., Salesky, E., Mielke, S. J., Lee, W. Y., Sharma, A., Santilli, A., Chaffin, A., Stiegler,

A., Datta, D., Szczechla, E., Chhablani, G., Wang, H., Pandey, H., Strobelt, H., Fries, J. A., Rozen, J., Gao, L., Sutawika, L., Bari, M. S., Al-shaibani, M. S., Manica, M., Nayak, N., Teehan, R., Albanie, S., Shen, S., Ben-David, S., Bach, S. H., Kim, T., Bers, T., Fevry, T., Neeraj, T., Thakker, U., Raunak, V., Tang, X., Yong, Z.-X., Sun, Z., Brody, S., Uri, Y., Tojarieh, H., Roberts, A., Chung, H. W., Tae, J., Phang, J., Press, O., Li, C., Narayanan, D., Bourfoune, H., Casper, J., Rasley, J., Ryabinin, M., Mishra, M., Zhang, M., Shoeybi, M., Peyrounette, M., Patry, N., Tazi, N., Sanseviero, O., von Platen, P., Cornette, P., Lavallée, P. F., Lacroix, R., Rajbhandari, S., Gandhi, S., Smith, S., Requena, S., Patil, S., Dettmers, T., Baruwa, A., Singh, A., Cheveleva, A., Ligozat, A.-L., Subramonian, A., Névéol, A., Lovering, C., Garrette, D., Tunuguntla, D., Reiter, E., Taktasheva, E., Voloshina, E., Bogdanov, E., Winata, G. I., Schoelkopf, H., Kalo, J.-C., Novikova, J., Forde, J. Z., Clive, J., Kasai, J., Kawamura, K., Hazan, L., Carpuat, M., Clinciu, M., Kim, N., Cheng, N., Serikov, O., Antverg, O., van der Wal, O., Zhang, R., Zhang, R., Gehrmann, S., Pais, S., Shavrina, T., Scialom, T., Yun, T., Limisiewicz, T., Rieser, V., Protasov, V., Mikhailov, V., Pruksachatkun, Y., Belinkov, Y., Bamberger, Z., Kasner, Z., Rueda, A., Pestana, A., Feizpour, A., Khan, A., Faranak, A., Santos, A., Hevia, A., Unldreaj, A., Aghagol, A., Abdollahi, A., Tammour, A., HajiHosseini, A., Behroozi, B., Ajibade, B., Saxena, B., Ferrandis, C. M., Contractor, D., Lansky, D., David, D., Kiela, D., Nguyen, D. A., Tan, E., Baylor, E., Ozoani, E., Mirza, F., Ononiwu, F., Rezanejad, H., Jones, H., Bhattacharya, I., Solaiman, I., Sedenko, I., Nejadgholi, I., Passmore, J., Seltzer, J., Sanz, J. B., Fort, K., Dutra, L., Samagaio, M., Elbadri, M., Mieskes, M., Gerchick, M., Akinlolu, M., McKenna, M., Qiu, M., Ghauri, M., Burynok, M., Abrar, N., Rajani, N., Elkott, N., Fahmy, N., Samuel, O., An, R., Kromann, R., Hao, R., Alizadeh, S., Shubber, S., Wang, S., Roy, S., Viguier, S., Le, T., Oyebade, T., Le, T., Yang, Y., Nguyen, Z., Kashyap, A. R., Palasciano, A., Callahan, A., Shukla, A., Miranda-Escalada, A., Singh, A., Beilharz, B., Wang, B., Brito, C., Zhou, C., Jain, C., Xu, C., Fourrier, C., Periñán, D. L., Molano, D., Yu, D., Manjavacas, E., Barth, F., Fuhrimann, F., Altay, G., Bayrak, G., Burns, G., Vrabec, H. U., Bello, I., Dash, I., Kang, J., Giorgi, J., Golde, J., Posada, J. D., Sivaraman, K. R., Bulchandani, L., Liu, L., Shinzato, L., de Bykhovetz, M. H., Takeuchi, M., Pàmies, M., Castillo, M. A., Nezhurina, M., Sänger, M., Samwald, M., Cullan, M., Weinberg, M., De Wolf, M., Mihaljcic, M., Liu, M., Freidank, M., Kang, M., Seelam, N., Dahlberg, N., Broad, N. M., Muellner, N., Fung, P., Haller, P., Chandrasekhar, R., Eisenberg, R., Martin, R., Canalli, R., Su, R., Su, R., Cahyawijaya, S., Garda, S., Deshmukh, S. S., Mishra, S., Kiblawi, S., Ott, S., Sangaroonsiri, S., Kumar, S., Schweter, S., Bharati, S., Laud, T., Gigant, T., Kainuma, T., Kusa, W., Labrak, Y., Bajaj, Y. S., Venkatraman, Y., Xu, Y., Xu, Y., Xu, Y., Tan, Z., Xie, Z., Ye, Z., Bras, M., Belkada, Y., and Wolf, T. Bloom: A 176b-parameter open-access multilingual language model. 2022. doi: 10.48550/ARXIV.2211.05100. URL https://arxiv.org/abs/2211.05100.

Lee, J., Le, T., Chen, J., and Lee, D. Do language models plagiarize? *arXiv preprint arXiv:2203.07618*, 2022.

Lee, K., Cooper, A. F., and Grimmelmann, J. Talkin"bout ai generation: Copyright and the generative-ai supply chain. *arXiv preprint arXiv:2309.08133*, 2023.

Li, J., Karamolegkou, A., Kementchedjhieva, Y., Abdou, M., Lehmann, S., and Søgaard, A. Structural similarities between language models and neural response measurements, 2023.

Liang, P., Bommasani, R., Creel, K., and Reich, R. The time is now to develop community norms for the release of foundation models. *City*, 2022a.

Liang, P., Bommasani, R., Creel, K. A., and Reich, R. The time is now to develop community norms for the release of foundation models, 2022b. URL https://crfm.stanford.edu/2022/05/17/community-norms.html.

Lipp, M., Schwarz, M., Gruss, D., Prescher, T., Haas, W., Mangard, S., Kocher, P., Genkin, D., Yarom, Y., and Hamburg, M. Meltdown, January 2018. URL http://arxiv.org/abs/1801.01207. arXiv:1801.01207 [cs].

Liu, D., Metzman, J., Chang, O., and Google Open Source Security Team. AI-Powered Fuzzing: Breaking the Bug Hunting Barrier, 2023. URL https://security.googleblog.com/2023/08/ai-powered-fuzzing-breaking-bug-hunting.html.

Llach, L. Naked deepfake images of teenage girls shock spanish town: But is it an ai crime? *EuroNews*, 2023. URL https://www.euronews.com/next/2023/09/24/spanish-teens-received-deepfake-ai-nudes-of-themselves-but-is-it-a-crime.

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *arXiv preprint arXiv:2310.16787*, 2023a.

Longpre, S., Yauney, G., Reif, E., Lee, K., Roberts, A., Zoph, B., Zhou, D., Wei, J., Robinson, K., Mimno, D., et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*, 2023b.

Lucas, F., Lofgren, Z., Collins, M., Obernolte, J., Stevens, H., and Foushee, V. Letter to the National Institute of Standards and Technology Director Laurie Locascio from the House Committee on Science, Space, and Technology, December 2023. URL https://republicans-science.house.gov/_cache/files/8/a/8a9f893d-858a-419f-9904-52163f22be71/191E586AF744B32E6831A248CD7F4D41.2023-12-14-aisi-scientific-merit-final-signed.pdf.

Maiberg, E. Civitai and OctoML Introduce Radical New Measures to Stop Abuse After 404 Media Investigation, December 2023a. URL https://www.404media.co/civitai-and-octoml-introduce-radical-new-measures-to-stop-abuse-after-404-media-investigation/.

Maiberg, E. Inside the AI Porn Marketplace Where Everything and Everyone Is for Sale, August 2023b. URL https://www.404media.co/inside-the-ai-porn-marketplace-where-everything-and-everyone-is-for-sale/.

Matthews, D. Scientists grapple with risk of artificial intelligence-created pandemics, 2023. URL https://sciencebusiness.net/news/ai/scientists-grapple-risk-artificial-intelligence-created-pandemics.

MITRE. Adversarial threat landscape for ai systems, 2021. URL https://atlas.mitre.org/.

Mortimer, S. StopNCII.org has launched, December 2021. URL https://revengepornhelpline.org.uk/news/stopncii-org-has-launched/.

Mouton, C. A., Lucas, C., and Guest, E. The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. Technical report, RAND Corporation, January 2024. URL https://www.rand.org/pubs/research_reports/RRA2977-2.html.

Mozilla. Joint Statement on AI Safety and Openness, October 2023. URL https://open.mozilla.org/letter/.

Musser, M. A Cost Analysis of Generative Language Models and Influence Operations, August 2023. URL http://arxiv.org/abs/2308.03740. arXiv:2308.03740 [cs].

Narayanan, A. and Kapoor, S. Generative ai companies must publish transparency reports, 2023a. URL https://knightcolumbia.org/blog/generative-ai-companies-must-publish-transparency-reports.

Narayanan, A. and Kapoor, S. GPT-4 and professional benchmarks: the wrong answer to the wrong question, March 2023b. URL https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks.

Narayanan, A., Kapoor, S., and Lazar, S. Model alignment protects against accidental harms, not intentional ones, December 2023. URL https://www.aisnakeoil.com/p/model-alignment-protects-against.

National Telecommunications and Information Administration. Dual use foundation artificial intelligence models with widely available model weights. Request for Comment, 02 2024. URL https://www.ntia.gov/federal-register-notice/2024/dual-use-foundation-artificial-intelligence-models-widely-available. Federal Register.

Nelson, A. et al. Comment of the ai policy and governance working group on the ntia request for comment on dual use foundation artificial intelligence models with widely available model weights. Docket NTIA-240216-0052, March 2024.

Noble, S. U. *Algorithms of Oppression*. New York University Press, 2018.

OpenAI. Better language models and their implications, 2019a. URL https://openai.com/research/better-language-models.

OpenAI. GPT-2: 1.5B release, 2019b. URL https://openai.com/research/gpt-2-1-5b-release.

Ovadya, A. and Whittlestone, J. Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning, July 2019. URL http://arxiv.org/abs/1907.11274. arXiv:1907.11274 [cs].

Padmakumar, V. and He, H. Does writing with language models reduce content diversity?, 2023.

Paris, B. and Donovan, J. Deepfakes and Cheap Fakes, September 2019. URL https://datasociety.net/library/deepfakes-and-cheap-fakes/. Publisher: Data & Society Research Institute.

Park, J. S., Popowski, L., Cai, C., Morris, M. R., Liang, P., and Bernstein, M. S. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, pp.

1–18, New York, NY, USA, October 2022. Association for Computing Machinery. ISBN 978-1-4503-9320-1. doi: 10.1145/3526113.3545616. URL https://dl.acm.org/doi/10.1145/3526113.3545616.

Patil, V., Hase, P., and Bansal, M. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. *arXiv preprint arXiv:2309.17410*, 2023.

Patwardhan, T., Liu, K., Markov, T., Chowdhury, N., Leet, D., Cone, N., Maltbie, C., Huizinga, J., Wainwright, C., Jackson, S. F., Adler, S., Casagrande, R., and Madry, A. Building an early warning system for LLM-aided biological threat creation, January 2024. URL https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation.

Perrigo, B. Openai used kenyan workers on less than 2 per hour to make chatgpt less toxic, Jan 2023. URL https://time.com/6247678/openai-chatgpt-kenya-workers/.

Pipatanakul, K., Jirabovonvisut, P., Manakul, P., Sripaisarnmongkol, S., Patomwong, R., Chokchainant, P., and Tharnpipitchai, K. Typhoon: Thai Large Language Models, December 2023. URL http://arxiv.org/abs/2312.13951. arXiv:2312.13951 [cs].

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to!, 2023.

Quintero, B. Introducing virustotal code insight: Empowering threat analysis with generative ai, apr 2023. URL https://blog.virustotal.com/2023/04/introducing-virustotal-code-insight.html.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019a.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language Models are Unsupervised Multitask Learners. 2019b. URL https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe.

Raffel, C. Building Machine Learning Models Like Open Source Software, February 2023. URL https://cacm.acm.org/magazines/2023/2/268952-building-machine-learning-models-like-open-source-software/fulltext.

Raji, I. D. and Buolamwini, J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pp. 429–435, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314244. URL https://doi.org/10.1145/3306618.3314244.

Reid, S. The Deepfake Dilemma: Reconciling Privacy and First Amendment Protections, June 2020. URL https://papers.ssrn.com/abstract=3636464.

Roberts, J. J. Fake Porn Videos Are Terrorizing Women. Do We Need a Law to Stop Them?, January 2019. URL https://fortune.com/2019/01/15/deepfakes-law/.

Saliba, E. Sharing deepfake pornography could soon be illegal in America, June 2023. URL https://abcnews.go.com/Politics/sharing-deepfake-pornography-illegal-america/story?id=99084399.

Sandbrink, J. B. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools, 2023.

Sastry, G. Beyond "release" vs. "not release", 2021. URL https://crfm.stanford.edu/commentary/2021/10/18/sastry.html.

Satter, R. FBI says artificial intelligence being used for 'sextortion' and harassment. *Reuters*, June 2023. URL https://www.reuters.com/world/us/fbi-says-artificial-intelligence-being-used-sextortion-harassment-2023-06-07/.

Scott, D. Deepfake Porn Nearly Ruined My Life, February 2020. URL https://www.elle.com/uk/life-and-culture/a30748079/deepfake-porn/. Section: Life + Culture.

Seaman, J. Cyber threat prediction and modelling. In Montasari, R. (ed.), *Artificial Intelligence and National Security*. Springer, Cham, 2022. doi: 10.1007/978-3-031-06709-9_7.

Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., Winter, C., Arnold, M., Ó hÉigeartaigh, S., Korinek, A., et al. Open-sourcing

highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. 2023.

Service, R. F. Could chatbots help devise the next pandemic virus?, 2023. URL https://www.science.org/content/article/could-chatbots-help-devise-next-pandemic-virus.

Shevlane, T. Structured access: an emerging paradigm for safe ai deployment. *arXiv preprint arXiv:2201.05159*, 2022.

Shevlane, T. and Dafoe, A. The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse? In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 173–179, New York, NY, USA, February 2020. Association for Computing Machinery. ISBN 978-1-4503-7110-0. doi: 10.1145/3375627.3375815. URL https://doi.org/10.1145/3375627.3375815.

Shostack, A. *Threat Modeling: Designing for Security*. John Wiley & Sons, Feb 2014. ISBN 978-1-118-80999-0. The only security book to be chosen as a Dr. Dobbs Jolt Award Finalist since Bruce Schneier's Secrets and Lies and Applied Cryptography!

Siddique, H. Sharing deepfake intimate images to be criminalised in England and Wales. *The Guardian*, June 2023. ISSN 0261-3077. URL https://www.theguardian.com/society/2023/jun/27/sharing-deepfake-intimate-images-to-be-criminalised-in-england-and-wales.

Soice, E. H., Rocha, R., Cordova, K., Specter, M., and Esvelt, K. M. Can large language models democratize access to dual-use biotechnology?, 2023.

Solaiman, I. The gradient of generative ai release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 111–122, 2023.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., and Wang, J. Release Strategies and the Social Impacts of Language Models, November 2019a. URL http://arxiv.org/abs/1908.09203. arXiv:1908.09203 [cs].

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., and Wang, J. Release strategies and the social impacts of language models. *ArXiv*, abs/1908.09203, 2019b.

Solsman, J. E. A deepfake bot is creating nudes out of regular photos, October 2020. URL https://www.cnet.com/news/privacy/deepfake-bot-on-telegram-is-violating-women-by-forging-nudes-from-regular-pics/.

Staff, A. Meltdown and Spectre: Here's what Intel, Apple, Microsoft, others are doing about it, January 2018. URL https://arstechnica.com/gadgets/2018/01/meltdown-and-spectre-heres-what-intel-apple-microsoft-others-are-doing-about-it/.

Sweeney, L. Discrimination in online ad delivery. *Queue*, 11(3):10:10–10:29, March 2013. ISSN 1542-7730. doi: 10.1145/2460276.2460278. URL http://doi.acm.org/10.1145/2460276.2460278.

Takanen, A., Demott, J. D., and Miller, C. *Fuzzing for Software Security*. Artech House Publishers, Norwood, MA, 1st edition edition, July 2008. ISBN 978-1-59693-214-2.

Thiel, D. Identifying and eliminating csam in generative ml training data and models. Technical report, Technical report, Stanford University, Palo Alto, CA, 2023. URL https://purl . . . , 2023.

Thiel, D. and Einstein, L. Online Consent Moderation, December 2020. URL https://cyber.fsi.stanford.edu/io/news/ncii-legislation-limitations.

Thiel, D., Stroebel, M., and Portnoff, R. Generative ML and CSAM: Implications and Mitigations. 2023. doi: 10.25740/jv206yg3793. URL https://purl.stanford.edu/jv206yg3793.

Tian, Y., Wang, Y., Zhang, Z., Chen, B., and Du, S. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. *arXiv preprint arXiv:2310.00535*, 2023.

Toma, A., Senkaiahliyan, S., Lawler, P. R., Rubin, B., and Wang, B. Generative AI could revolutionize health care — but not if control is ceded to big tech. *Nature*, 624(7990):36–38, December 2023. doi: 10.1038/d41586-023-03803-y. URL https://www.nature.com/articles/d41586-023-03803-y. Bandiera_abtest: a Cg_type: Comment Number: 7990 Publisher: Nature Publishing Group Subject_term: Machine learning, Health care, Medical research, Technology.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

UK AISI. Introducing the ai safety institute, 2023. URL https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.

UK CMA. Ai foundation models: Initial report, 2023. URL https://assets.publishing.service.gov.uk/media/65081d3aa41cc300145612c0/Full_report_.pdf.

Vincent, J. Meta's powerful AI language model has leaked online — what happens now?, March 2023. URL https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse.

Vinsel, L. *Moving violations: automobiles, experts, and regulations in the United States*. JHU Press, 2019.

Vipra, J. and Korinek, A. Market concentration implications of foundation models: The invisible hand of chatgpt. *The Brookings Institution*, 2023. URL https://www.brookings.edu/articles/market-concentration-implications-of-foundation-models-the-invisible-hand-of-chatgpt.

Wang, A. and Russakovsky, O. Directional Bias Amplification. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 10882–10893. PMLR, July 2021. URL https://proceedings.mlr.press/v139/wang21t.html. ISSN: 2640-3498.

Wang, B. and Komatsuzaki, A. Gpt-j-6b: A 6 billion parameter autoregressive language model, 2021.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *The Eleventh International Conference on Learning Representations*, 2022.

Widder, D. G., Nafus, D., Dabbish, L., and Herbsleb, J. Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 2035–2046, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533779. URL https://dl.acm.org/doi/10.1145/3531146.3533779.

Widder, D. G., West, S., and Whittaker, M. Open (for business): Big tech, concentrated power, and the political economy of open ai. 2023.

Xia, M., Gao, T., Zeng, Z., and Chen, D. Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning, October 2023. URL http://arxiv.org/abs/2310.06694. arXiv:2310.06694 [cs].

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

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. Defending against neural fake news. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9054–9065, 2019.

Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., and Qiao, Y. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention, June 2023. URL http://arxiv.org/abs/2303.16199. arXiv:2303.16199 [cs].

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. OPT: Open pre-trained transformer language models. *arXiv*, 2022.

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

Paper	Domain	Summary of Research
Yang et al. (2023)	Safety and security	Safety alignment can be subverted with minimal finetuning.
Choi et al. (2023)	Privacy and security	Assessing the ability to verify a model's training data.
Patil et al. (2023)	Privacy and security	Methods to prevent sensitive information extraction attacks.
Zou et al. (2023)	Safety and alignment	Adversarial attacks that transfer from open models to black-box, closed models.
Kirchenbauer et al. (2023)	Content provenance	Watermarking methods for LLMs
Dettmers et al. (2023)	Training efficiency	Efficient training with quantized low rank adapters.
Longpre et al. (2023b)	Toxicity and bias	Understanding the effects of "quality" filters on model toxicity and performance.
Li et al. (2023)	Brain imaging analysis	Comparing the representations of sequences in LLMs and neural response measurements.
Han et al. (2023)	Medical applications	Training medical application models.
Wang et al. (2022)	Interpretability	Explaining model behaviors in terms of their internal components.
Tian et al. (2023)	Architecture analysis	Understanding training dynamics in multi-layer Transformer architectures.

Table A1. A non-comprehensive list of research that uses open foundation models, organized by the research domain. The example works are enabled with GPT-2 (Radford et al., 2019a), LLaMA (Touvron et al., 2023a), Llama-2 (Touvron et al., 2023b), Pythia (Biderman et al., 2023), GPT-J (Wang & Komatsuzaki, 2021), GPT-NeoX (Black et al., 2022), Bloom (Le Scao et al., 2022), and OPT (Zhang et al., 2022). Based on data compiled by Biderman (2023).

Appendix

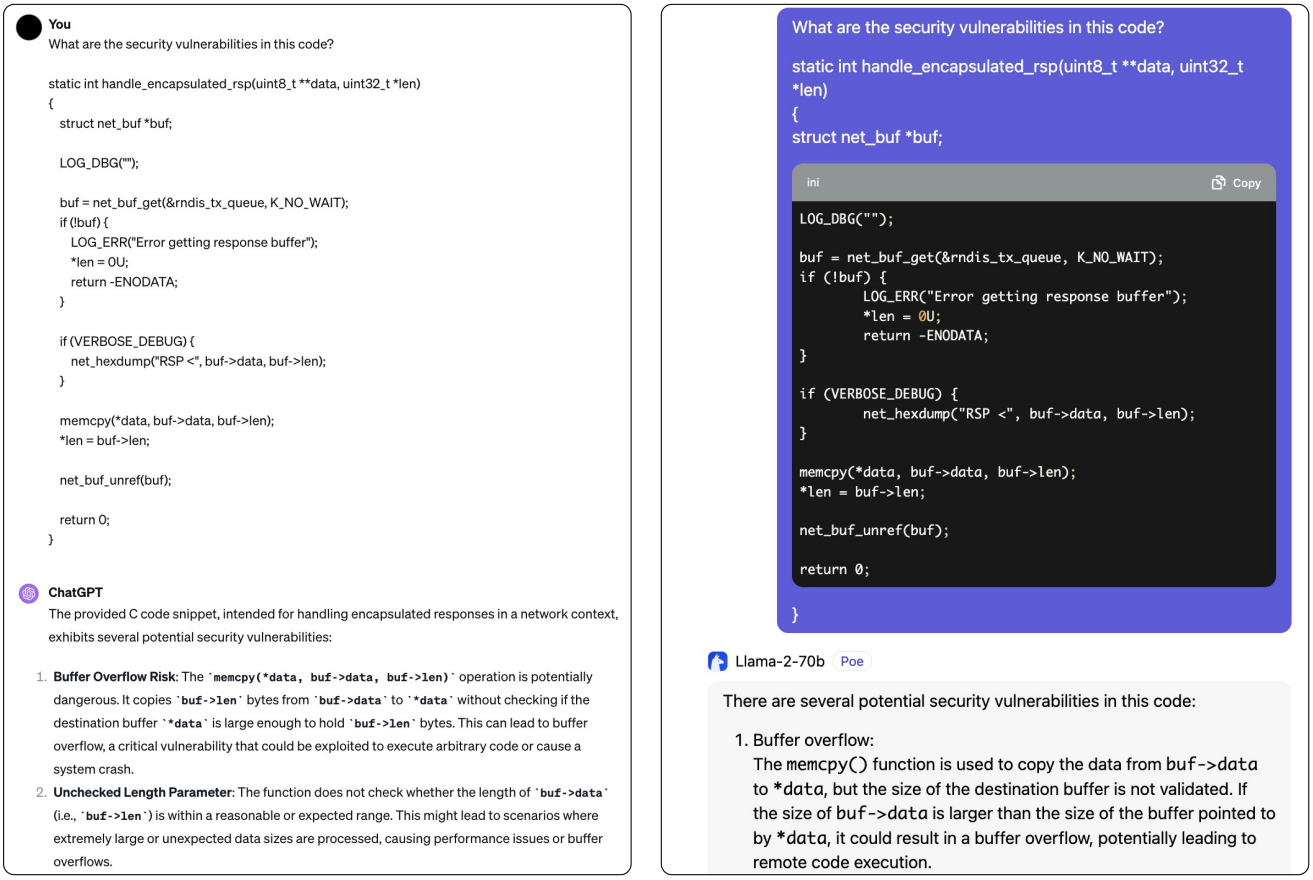
A. A brief history of open foundation models

In the 2010s, the advent of sophisticated generative image models introduced the concern of deepfakes that could misinform viewers (Paris & Donovan, 2019; Chesney & Citron, 2018; Widder et al., 2022; Cole, 2017). These models facilitated significant impersonation, political misinformation, and non-consensual intimate imagery (NCII). For example, a Telegram bot was used to generate over a hundred thousand nude images of women (Solsman, 2020). These models were not foundation models; they relied on more rudimentary algorithms, such as swapping faces in images.

In the late 2010s, foundation models gave rise to powerful generative capabilities for language (Radford et al., 2018), and new misuse concerns came to the fore. In February 2019, OpenAI announced the GPT-2 series of models (Radford et al., 2019b; OpenAI, 2019a): four language models ranging between 124 million and 1.5 billion parameters. Primed by several concerns — especially the potential for large-scale disinformation (Solaiman et al., 2019a) — OpenAI opted for a staged release where models of increasing size were openly released from February to November 2019. While the company did not ultimately find evidence of model misuse during the staged release (OpenAI, 2019b), this process brought attention to the nexus between release and misuse (Bommasani et al., 2021; Sastry, 2021; Shevlane, 2022; Liang et al., 2022b; Solaiman, 2023).

Since the release of GPT-2, hundreds of foundation models have been released by various actors adopting different release strategies (Bommasani et al., 2023c). The open release of Stable Diffusion in August 2022 was particularly salient, as it was one of the first text-to-image models to be available widely outside the research community. However, because the model weights were shared publicly, users easily circumvented filters that Stability AI implemented to prevent the generation of not safe for work (NSFW) imagery. As a result, AI-generated pornography based on Stable Diffusion offshoots quickly spread across the internet, including images resembling real people generated without their consent (Maiberg, 2023b).

Meta's release of its LLaMA language models (Touvron et al., 2023a) in March 2023 marked another significant event in the trajectory of open foundation models. LLaMA was released via a form that allowed researchers to download model weights by accepting a license for non-commercial use. But the model weights were quickly leaked, leading to concerns that the highly capable model could facilitate misuse. U.S. Senators Hawley and Blumenthal sent a letter to Meta CEO Mark Zuckerberg expressing concerns about Meta's release strategy (Blumenthal & Hawley, 2023a).



(a) When prompted with the code from a recent security vulnerability (Flavio Ceolin, 2023), ChatGPT is able to diagnose the issue.

(b) Llama-2 is similarly able to spot the buffer overflow issue.

Figure A1. Comparison of ChatGPT and Llama-2 in identifying vulnerabilities

B. Risk assessment framework

We surveyed the literature on misuse vectors specifically associated with open foundation models, identifying biosecurity, cybersecurity, voice cloning scams, spear phishing, disinformation, non-consensual intimate imagery, and child sexual abuse material. For each of these misuse vectors, we chose a past study that analyzed the misuse vector and discussed or analyzed openness. Note that these studies did not need to compare open models to closed ones or other existing technologies, such as the internet, to be considered. And the lack of evidence in one of the risk assessment framework elements does not mean that the study's analysis is flawed; it only means that the study does not present adequate evidence of an increased marginal societal risk of open foundation models. Below, we present our justifications for the scores for each study in Table 1.

B.1. Spear-phishing scams (Hazell, 2023)

Threat identification ● The threat is specified clearly—scams due to emails that appear personalized. The threat actor is specified as low-skill actors who can focus on high-level planning and outsource the email writing to LLMs. (See sections 4 and 5.)

Existing risk (absent open FMs) ◑ There is some characterization of existing risk (e.g., examples of successful phishing scams), but no broader analysis of how societally impactful spear-phishing scams currently are. This is essential to understanding the arguments' scope of overall and marginal risk. (See sections 2 and 4.)

Existing defenses (absent open FMs) ○ While there is a brief discussion of some defenses, there is little analysis of how effective protections such as those built into modern operating systems or email services are at preventing such spear-phishing scams. (See, for example, section 7.2.)

Evidence of marginal risk ○ There is no analysis of how open FMs can be used to circumvent existing defenses built into operating systems or email services. (See, for example, section 5.3.).

Ease of defense ◑ The paper discusses AI-based defenses (e.g., using Sec-PaLM), and discusses how foundation models should be governed given the findings of the study, which may provide a further defense. However, there is no analysis on how robust existing defenses are, and whether further defenses are needed. (See section 7.)

Uncertainty/assumptions ○ The paper does not explicitly analyze assumptions underlying the analysis (e.g., existing defenses will fail; content creation is already cheap without AI models) or how these assumptions could fail in the real world. (See, for example, section 5.)

B.2. Cybersecurity risk (Seger et al., 2023)

Seger et al. (2023) outline several risks of releasing foundation models openly (including many of the other risks we look at—disinformation, scams etc.); here, we focus on their analysis of cybersecurity risks.

Threat identification ◑ The main threat model described is the creation of malware, in very broad terms. No clear threat actors are identified. (See section 3.1.1.)

Existing risk (absent open FMs) ○ The paper does not discuss how prevalent malware-based cyberattacks currently are, or their societal impact, or provide examples of the type of cyberattacks that are in scope. (See, for example, section 3.1.1.)

Existing defenses (absent open FMs) ◑ Some examples of defenses (e.g., bug bounties) are provided, but there is no in-depth analysis of how helpful these defenses are against existing risks, excluding AI systems. (See section 4.1.3.)

Evidence of marginal risk ○ The paper does not provide any evidence that open foundation models have contributed to cybersecurity incidents by facilitating the creation of malware. It does not compare against other baselines, such as finding similar information on the Internet or using closed models (e.g., via jailbreaks or fine tuning). (See, for example, sections 3.1 and 3.2.)

Ease of defense ◑ The paper briefly discusses bug bounties as a mechanism for hiring experts to bolster defense, but does not analyse how much interventions reduce marginal risk. It also acknowledges the arguments of open models advancing safety research, but does not assess alternative ways to mitigate marginal risk (aside from staged release) such as using LLMs as tools for cyberdefense. (See section 4.1.3.)

Uncertainty/assumptions ○ The paper acknowledges that the offense-defense balance is tentative, but does not analyze it in the context of cybersecurity or articulate other core assumptions that affect the assessment of open and closed models, such as the resources available to the most concerning threat actors. (See, for example, sections 3.1.1 and 3.1.3.)

B.3. Disinformation (Musser, 2023)

Threat identification ● The threat is identified clearly: propagandists can cheaply create content to conduct influence operations. (See sections 1 and 2.)

Existing risk (absent open FMs) ◑ The paper gives some examples of preexisting risks of mass content generation (e.g., by the Chinese government and Russian agencies). It does not, however, analyze the societal impact of such influence operations beyond the number of content posts generated, or the effectiveness of current methods at influencing people at scale. (See section 2.)

Existing defenses (absent open FMs) ○ The paper gives makes some reference to efforts by social media companies and language model developers to defend against influence operations, but it does not analyze the effectiveness of these or other existing defenses. (See sections 3 and 4.)

Evidence of marginal risk ○ The paper gives evidence of cost reduction for the generation of content, but does not discuss existing assessments of the demand for such content, whether it is actually effective in the real world, or if open foundation models can be used to circumvent existing defenses. It also does not analyze the costs associated with distributing said content (e.g., maintaining user profiles, avoiding behavioral social media moderation etc.), which is likely to overshadow the cost of content creation in disinformation operations. (See, for example, sections 3 and 4.)

Ease of defense ◑ The paper analyzes the cost of avoiding monitoring on closed models compared to open ones. It does not analyze steps social media platforms could take such as behavioral or network-based moderation, improving captcha to

detect bots, or other existing mechanisms already in use by social media platforms. (See section 5.)

Uncertainty/assumptions ● The paper articulates several assumptions built into the analysis, such as the type of actors using the models, the relative costs of open vs. closed models, and the cost of evading monitoring for closed foundation models. (See sections 4, 5 and 8.)

B.4. Biosecurity risk (Gopal et al., 2023)

Threat identification ● The threat is identified clearly: individuals without training could obtain and create pandemic-causing pathogens. (See page 2.)

Existing risk (absent open FMs) ○ The paper provides no analysis of how a rogue individual or non-state actor could engage in bioterrorism absent foundation models (or open foundation models). (See, for example, page 2.)

Existing defenses (absent open FMs) ◑ There is a brief analysis of limitations to rogue individuals' and non-state actors' ability to leverage information about materializing biorisks. The paper mentions deterring effects such as the lack of information about viruses, the number of researchers capable of assembling an influenza virus, and existing immunity from historical pathogens. But there is no mention of other existing notable defenses such as controls on procurement of raw materials or benchtop DNA synthesizers. Nor is there any discussion of the efficacy of these defenses. (See page 3.)

Evidence of marginal risk ○ The paper gives no comparison to similar risks based on widely available information on the Internet. The discussion of closed models' safeguards is not substantiated by evidence; for example, the paper does not analyze the risks from closed models related to jailbreaking or fine tuning. (See, for example, page 6.)

Ease of defense ○ The paper proposes defenses such as legal liability and market-based insurance. There is no evidence regarding the feasibility or effectiveness of these proposed defenses, especially against well resourced rogue actors. Similarly, there is no analysis of how effective existing defenses (such as controls on procurement) would be. (See, for example, pages 7 and 8.)

Uncertainty/assumptions ○ The paper does not analyze assumptions built into the analysis, such as how future capabilities would evolve; whether current models are already capable enough to be fine tuned on biological information and generate information on how to cause *novel* pandemics (which increases marginal risk); or how information on the Internet could similarly aid attackers (which would lower marginal risk). (See, for example, page 7.)

B.5. Voice cloning (Ovadya & Whittlestone, 2019)

Threat identification ● The threat is identified clearly: financial scams due to voice cloning technology using machine learning. The paper identifies several potential threat actors, such as malicious users with or without ML expertise. (See section 2.2.)

Existing risk (absent open FMs) ◑ There is some characterization of existing risk (such as financial crimes), but there is little broader analysis of the societal impact of voice cloning scams at present. (See sections 1 and 3.5.)

Existing defenses (absent open FMs) ◑ There is some discussion of existing defenses implemented by companies (Synthesia, Lyrebird), but little analysis of their efficacy. There is no analysis of the efficacy of these defenses against preexisting risk. (See section 3.5.)

Evidence of marginal risk ◑ There is some discussion of the marginal risk of openness (e.g., related to reproducibility, modifiability, access ratchets), but minimal direct comparison to risks associated with closed models or other technologies that could similarly enable misuse. (See sections 2.2 and 4)

Ease of defense ◑ The paper analyzes several mechanisms for mitigating marginal risk (e.g. timing of release, what assets are released) but does not give a thorough account of how much these mitigations could address marginal risk. (See sections 3.3 and 3.4.)

Uncertainty/assumptions ● The paper articulates several assumptions built into its analysis such as the current state of open and closed models and how they are released. (See sections 2.2, 3.5, and 4.2-4.4.)

B.6. Non-consensual intimate imagery (NCII) (Lakatos, 2023)

Threat identification ● The threat is identified clearly: digitally altered NCII. Threat actors are people who do not necessarily require machine learning skills and can rely on easy-to-use interfaces in order to create NCII. (See page 1.)

Existing risk (absent open FMs) ◑ The paper looks at comments referring users to NCII distribution sites before (and after) the widespread availability of open text-to-image foundation models and demonstrates that there are tens of millions of unique visitors to such sites, though it does not describe the risk of preexisting tools such as those for face swapping. (See pages 1 and 3.)

Existing defenses (absent open FMs) ○ The paper does not examine existing technical or legal defenses against NCII that could be used to deter malicious users or prevent the spread of NCII. (See, for example, page 1.)

Evidence of marginal risk ◑ The paper outlines the risk stemming from open models, such as by showing that the increase in use of services for creating NCII is driven by open FMs, but does not compare the risk of open models to that of closed models or other digital technologies such as Photoshop. (See page 1.)

Ease of defense ◑ There is some discussion of downstream entities such as Paypal not providing services to platforms that enable the distribution of NCII, but there is no significant analysis of defenses against marginal risk posed by open foundation models specifically. (See page 5.)

Uncertainty/assumptions ○ There is no explicit discussion of the uncertainty or assumptions built into the analysis, such as the potential adaptation of defenses. (See, for example, page 1.)

B.7. Child sexual abuse material (CSAM) (Thiel et al., 2023)

Threat identification ● The threat is clearly identified: the distribution of computer-generated CSAM (or CG-CSAM). It identifies threat actors as hobbyist groups and forums for sharing CSAM. (See section 1.)

Existing risk (absent open FMs) ● The authors provide evidence that the existing prevalence of CG-CSAM is low. Based on an analysis of online forums dedicated to sharing CSAM, they find that less than 1% of CSAM on online forums is photorealistic CG-CSAM. (See section 1.)

Existing defenses (absent open FMs) ● The paper outlines legal defenses as well as moderation by platforms (such as creating hashsets of known CSAM) as defenses against CSAM absent open FMs. (See sections 4 and 5.)

Evidence of marginal risk ● The authors list several concerns from the use of open foundation models: re-victimization of abused children if models are trained on their likeness to create new images; liar's dividend—when perpetrators who possess real CSAM claim they possess CG-CSAM; increasing toll on moderators who need to filter through increasingly lifelike CG-CSAM to find examples of real abused children; and the increased burden on enforcement agencies from a surge of CG-CSAM. The paper also provides evidence that most of CG-CSAM is based on the Stable Diffusion series of models, which are openly available; points out that watermarking or monitoring the use of Stable Diffusion is hard because of open weights; and discusses how such watermarking has been disabled. (See sections 4 and 5.)

Ease of defense ● The paper discusses several defenses (and analyses their shortcomings), including: developers removing known CSAM from training data; methods to create persistent watermarks that increase barriers to entry (such as Stable signature); legal defenses against photo-realistic CSAM; and low-tech defenses such as using EXIF data to differentiate CG-CSAM. (See sections 4, 5 and 6.)

Uncertainty/assumptions ● The paper points out the scope for better models in increasing risk; the assumption about the difficulty of watermarking open models; and the uncertainty about the legal status of CG-CSAM. (See sections 3, 4 and 5.)