Chapter 5

Hosting: Moderating Third-Party Models

Michael Veale and Robert Gorwa

Chapter 5 of *Malice, Models and Middlemen: The Full Stack Fight Against AI Misuse*, pages 116–146

Abstract

There are now many hosting platforms which allow lay or professional users to access, download and sometimes also tweak and manipulate various AI models. Since generative AI systems easily produce both illicit and legitimate material, and because hosts generally do not have visibility into the content generated by those systems (which can be downloaded and run locally or deployed in a cloud environment), it is difficult for hosts to effectively deal with the plethora of ways in which they are implicated in AI-powered intimate image abuse and other thriving misuse ecosystems. We begin by tracing the history of AI hosting platforms from early software development repositories to today's big players like Hugging Face, examining the legal and technical challenges facing hosting platforms seeking to do moderation in good faith. We then offer case studies demonstrating the distinct challenges facing three different types of hosting platform: general-purpose AI hosts like Hugging Face, specialized image generation platforms like Civitai and app stores increasingly needing to deal with AI apps or apps with AI integrations. While this layer of the stack is characterized by some wicked challenges (we focus in particular on the *dual-use nature* of AI models, the *imperfect information* that platforms need to deal with when making judgments about specific models in a sea of content and the ability of motivated actors to *obfuscate* misuse capabilities), we conclude by outlining the ample low-hanging fruit that could be picked if host actors brought on the requisite capacity and expertise to proactively deal with the most obvious forms of abuse.

Preprint, November 2026. Forthcoming from MIT Press, 2027. Page numbers follow the preprint of the full book.

Cite as: Michael Veale and Robert Gorwa, *Malice, Models and Middlemen: The Full Stack Fight Against AI Misuse* (MIT Press forthcoming 2027) ch 5

https://homepages.ucl.ac.uk/~ucqnmve/malicemodelsmiddlemen/chapter/5/

Copyright © 2026 Michael Veale and Robert Gorwa. All rights reserved. This page will be disabled for a period following publication of the physical book.

Reservation. The rightholder expressly reserves all rights to use this work for text and data mining, including under Article 4(3) of Directive (EU) 2019/790 and national implementations. This covers training, fine-tuning or evaluating AI models, and indexing or retrieval for AI systems.

Licensing. No licence is granted for these uses except as stated here.

Retrieval licence. Live retrieval to answer an individual user query is licensed only if the output shows visible credit and a working link to the source, quotes no more than 30 words, and doesn't replace reading the work. The content must not be used for training.

Other rights. Nothing here concedes that any unlicensed use is fair use or fair dealing.

5 HOSTING: MODERATING THIRD-PARTY MODELS

There are now many hosting platforms which allow lay or professional users to access, download and sometimes also tweak and manipulate various AI models. Since generative AI systems easily produce both illicit and legitimate material, and because hosts generally do not have visibility into the content generated by those systems (which can be downloaded and run locally or deployed in a cloud environment), it is difficult for hosts to effectively deal with the plethora of ways in which they are implicated in AI-powered intimate image abuse and other thriving misuse ecosystems. We begin by tracing the history of AI hosting platforms from early software development repositories to today's big players like Hugging Face, examining the legal and technical challenges facing hosting platforms seeking to do moderation in good faith. We then offer case studies demonstrating the distinct challenges facing three different types of hosting platform: general-purpose AI hosts like Hugging Face, specialized image generation platforms like Civitai and app stores increasingly needing to deal with AI apps or apps with AI integrations. While this layer of the stack is characterized by some wicked challenges (we focus in particular on the *dual-use nature* of AI models, the *imperfect information* that platforms need to deal with when making judgments about specific models in a sea of content and the ability of motivated actors to *obfuscate* misuse capabilities), we conclude by outlining the ample low-hanging fruit that could be picked if host actors brought on the requisite capacity and expertise to proactively deal with the most obvious forms of abuse.

TO GIT, OR NOT TO GIT

In late 2017, a Reddit user operating under the pseudonym 'deepfakes' began using what was then the cutting edge of open-source AI image-generation technology to create pornography.[1] Harnessing Google's open-source TensorFlow library to train neural networks with selected images obtained via searches, stock photos and YouTube videos, the user posted hardcore porn videos with grafted-in celebrity faces, 'featuring' Scarlett Johansson, Maisie Williams, Taylor Swift, Aubrey Plaza and Gal Gadot.[2] In December 2017, an account also using the 'deepfakes' handle created a public repository on the popular open-source software development platform GitHub. The project, called faceswap, could synthetically and (relatively) realistically overlay a face taken from one video onto another. In the months that followed, various similar projects appeared online, such as DeepFaceLab in summer 2018,[3] and DeepNude, software released in 2019 that "swaps clothes for naked breasts and a vulva, and only works on images of women."[4]

1: Samantha Cole, "AI-Assisted Fake Porn Is Here and We're All Fucked" (*Vice Motherboard*, 11 December 2017) https://web.archive.org/web/20180125040622/https://motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn accessed 11 January 2026; Samantha Cole, *How Sex Changed the Internet and the Internet Changed Sex* (Workman Publishing 2022) ch 15.

2: Cole, "AI-Assisted Fake Porn Is Here and We're All Fucked" (n 1).

3: This is a pipeline optimized for Nvidia GPUs, see https://github.com/iperov/DeepFaceLab.

4: Samantha Cole, "This Horrifying App Undresses a Photo of Any Woman With a Single Click" (*Vice Motherboard*, June 2019) https://www.vice.com/en/article/kzm59x/deepnude-app-creates-fake-nudes-of-any-woman accessed 16 July 2023.

These codebases could and would be used for everything from political satire to intentionally misleading forms of impersonation and disinformation. And the creators of these tools sometimes sought to make this clear, publicly disclaiming their use for harassment or non-consensual image generation.[5] The developers of the faceswap repository included an informal license as a README.md file on their project: "Faceswap is not for creating inappropriate content [for] any illicit, unethical, or questionable purposes."[6]

How would GitHub respond? Some repositories — with creators who may have been genuinely taken aback by the ways their tools were being used — were taken down voluntarily by the original creators after facing pushback from their peers.[7] Nevertheless, several copycat versions were quickly created. The core DeepNude model — based on a UC Berkeley research project — was reverse engineered and placed on GitHub again (albeit without the previously provided user interface).[8] But GitHub removed many other offending models — with policy staff noting that they "do not condone using GitHub for posting sexually obscene content and prohibit such conduct in our Terms of Service and Community Guidelines."[9] At the time, the relevant part of GitHub's community guidelines stated:

> Don't post content that is pornographic. This does not mean that all nudity, or all code and content related to sexuality, is prohibited. We recognize that sexuality is a part of life and non-pornographic sexual content may be a part of your project, or may be presented for educational or artistic purposes. We do not allow obscene sexual content or content that may involve the exploitation or sexualization of minors.[10]

While this policy may seem promising, it actually skirted around the core challenge facing the company. Enforcing rules around acceptable forms of text, images and videos to publish online had for years given platform companies 'content moderation' headaches,[11] and GitHub seemed to be no exception. Yet should AI models really be understood as 'content' in that same way? Or, given that they can be used to do things in the world that traditional content does not or cannot, are they something quite different? Given this, how should a platform like GitHub gauge what those models could or would actually do in the world?

5: Researchers have shown that despite these claims, sexual motives and typically non-consensual dynamics drove the development of this ecosystem. Aya Yadlin-Segal and Yael Oppenheim, "Whose Dystopia Is It Anyway? Deepfakes and Social Media Regulation" (2021) 27(1) Convergence 36 DOI: 10.1177/1354856520923963 ; Rachel Winter and Anastasia Salter, "DeepFakes: Uncovering Hardcore Open Source on GitHub" (2020) 7(4) Porn Studies 382 DOI: 10.1080/23268743.2019.1642794 .

6: deepfakes_faceswap, "GitHub Repository" (2023) https://github.com/deepfakes/faceswap/ accessed 30 October 2023.

7: Samantha Cole, "Creator of DeepNude, App That Undresses Photos of Women, Takes It Offline" (*Vice Motherboard*, June 2019) https://www.vice.com/en/article/qv7agw/deepnude-app-that-undresses-photos-of-women-takes-it-offline accessed 16 July 2023.

8: The model was a generative adversarial network (GAN) based on UC Berkeley's pix2pix system, see Katyanna Quach, "DeepNude's Makers Tried to Deep-Six Their Pervy AI App. Web Creeps Have Other Ideas: Cracked Copies Shared Online as Code Decompiled" (*The Register*, July 2019) https://www.theregister.com/2019/07/02/deepnude_ai_spreads/ accessed 16 July 2023; Phillip Isola and others, "Image-To-Image Translation With Conditional Adversarial Networks" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017).

9: Joseph Cox, "GitHub Removed Open Source Versions of DeepNude" (*Vice Motherboard*, July 2019) https://www.vice.com/en/article/8xzjpk/github-removed-open-source-versions-of-deepnude-app-deepfakes accessed 16 July 2023.

10: GitHub, "GitHub Community Guidelines - GitHub Help" (*GitHub Help (Internet Archive)*, June 2019) https://web.archive.org/web/20190616104903/https://help.github.com/en/articles/github-community-guidelines accessed 16 July 2023.

11: Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018); Robert Gorwa, "What Is Platform Governance?" (2019) 22(6) Information, Communication & Society 854 DOI: 10.1080/1369118X.2019.1573914 ; Robyn Caplan, "Networked Platform Governance: The Construction of the Democratic Platform" (2023) 17(0) International Journal of Communication 3451.

Models explicitly advertised by their creators as tools for creating non-consensual imagery, such as DeepNude, were taken down by GitHub. But faceswap and its clones posed a more difficult issue. They were described by their creators in generic, innocuous terms. In 2018, GitHub appeared to have gated the faceswap repository slightly, making it available for download to logged-in users only, although this was lifted in 2019.[12] The matter hasn't been resolved since then: as of early 2026, faceswap and many more advanced models for doing 'deepfake'-style image manipulation are easily available on GitHub to download.

This is not just a GitHub problem. There are now many hosting platforms where users can upload AI systems and AI-related datasets, which in turn can typically be downloaded and, depending on the business model, queried, tweaked or built upon by other users. Some of these services, like the one operated by New York-based firm Hugging Face, have become integrated into the daily activities of many legitimate researchers, startups and public organizations. They simultaneously host many thousands of open models, many of which can be, and in some cases are, used for criminal or malicious ends.

Furthermore, as transformer-powered text and image-generation models have become more powerful and also more lightweight, they are increasingly being integrated into mobile apps. This has put app stores into a similar position, where even to follow their current policies, they may need to be able to evaluate the capabilities of particularly complex and open-ended software. There is already evidence of generic, innocuous AI-image editing apps available in the Apple App Store simultaneously being advertised in shadier corners of the Web as specialty apps particularly optimized for non-consensual imagery generation and 'nudification.'[13]

General-purpose AI chatbots were implicated too, being distributed by app stores both directly and indirectly. This trend came to a head at the end of 2025, as xAI's Grok model, built into social media platform X (and also available as a standalone mobile app), was widely being used to synthetically undress celebrities, including minors.[14] The scandal triggered international action, with the UK government announcing a planned legal change to make supplying such apps a criminal offense.[15] But the government quickly realized that this might catch virtually all image and video manipulators, AI-enabled or not, and that it hadn't fully grasped the challenge posed by dual-use AI tools like Grok (which can deliver innocuous information as well as racist screeds and illegal non-consensual intimate image abuse).[16] The Secretary of State backtracked, confirming to the House of Commons Science and Technology Committee that the ban 'will apply to applications that have *one despicable purpose only*.'[17]

12: https://github.com/deepfakes/faceswap/issues/392. See the restriction active in the Internet Archive version at https://web.archive.org/web/20190203131424/https://github. com/login?return_to=https%3A%2F%2Fgithub.com%2Fdeepfakes%2Ffaceswap. Internet Archive versions later in 2019 are stored without a login wall, indicating that the limitation was lifted at some point then.

13: Emanuel Maiberg, "Apple Removes Nonconsensual AI Nude Apps Following 404 Media Investigation" (*404 Media*, 26 April 2024) https://www.404media.co/apple-removes-nonconsensual-ai-nude-apps-following-404-media-investigation/ accessed 3 November 2024.

14: AJ Vicens and Raphael Satter, "Elon Musk's Grok AI Floods X with Sexualized Photos of Women and Minors" (*Reuters*, 2 January 2026) https://www.reuters.com/legal/litigation/grok-says-safeguard-lapses-led-images-minors-minimal-clothing-x-2026-01-02/ accessed 13 February 2026.

15: HC Deb 12 January 2026, vol 778, col 607.

16: Lisa Hagen, "The Grok chatbot spews racist and antisemitic content" (*NPR*, 9 July 2025) https://www.npr.org/2025/07/09/nx-s1-5462609/grok-elon-musk-antisemitic-racist-content accessed 27 February 2026.

17: Liz Kendall, "Letter Entitled 'Intimate Grok AI Deepfakes' to Dame Chi Onwurah MP" (*House of Common Science and Technology Committee*, 12 January 2026) https://committees.parliament.uk/publications/51067/documents/283193/default/

As we will discuss, the opaque, complex and dual-use nature of many AI systems presents the companies that are hosting AI models in various forms with a wicked problem. Nonetheless, by sitting in between many open models and malign individuals and organizations seeking to deploy them, hosts are a little-discussed intermediary in the AI stack that could deploy policy interventions to disrupt some of the most obvious forms of misuse.

FROM SOFTWARE SHARING TO MODEL MARKETPLACES

Programmers, hackers, tinkerers and all other sorts of computer hobbyists have always shared and exchanged software. Long before the Web, software was being emailed and shared on bulletin board systems (BBSs). Magazine ads for modems in the late 1970s promised users the ability to 'swap programs with computer people everywhere.'[18] Usenet groups, notably *comp.sources.unix*, developed tools that allowed users to upload *patches* to software as threaded replies to facilitate collaboration and development. As collaborative tooling improved, and the Web emerged throughout the nineties, more polished software repositories were developed, such as the popular website SourceForge (launched in 1999). These were followed almost a decade later by GitHub, a platform based around the *git* tool created by Linux creator Linus Torvalds. GitHub fast became so central and popular in contemporary software development that Microsoft acquired it as a strategic asset in 2018.

As hype around 'big data' and machine learning grew throughout the 2010s, a few firms tried to create software platforms for algorithms. Algorithmia, a Seattle firm founded in 2013 by former Microsoft engineers, allowed third-party developers to upload algorithms "designed to slot right into new services."[19] It claimed it had thousands of algorithms and tens of thousands of enrolled developers, although it faded from view following its acquisition in the early 2020s.[20] Microsoft offered a similar service around the same time called the Cortana Intelligence Gallery. Yet unlike previous iterations of software repositories, the business model of these services relied not on open-source models, but rather on matchmaking, bringing firms providing specific tools together with clients with highly specific needs, and facilitating the smooth and sustainable integration of proprietary algorithms. It sat within the emerging practice of 'MLOps' — machine learning operations — which grew out of the notion of 'DevOps' (or development operations) in software engineering.[21] Yet these services remained small-scale and did not command very much commercial attention or interest.

accessed 13 February 2026.

18: Kevin Driscoll, *The Modem World: A Prehistory of Social Media* (Yale University Press 2022) 97.

19: Davey Alba, "Need Some AI? Yeah, There's a Marketplace for That" (*Wired*, September 2016) https://www.wired.com/2016/09/algorithmia-deep-learning/ accessed 28 August 2023.

20: Amit Chowdhry, "How Algorithmia Built The Largest Marketplace For Algorithms In The World" (*Forbes*, January 2018) https://www.forbes.com/sites/amitchowdhry/2018/01/22/how-algorithmia-built-the-largest-marketplace-for-algorithms-in-the-world/ accessed 14 September 2023; Taylor Soper, "DataRobot Lands $300M and Acquires Seattle Machine Learning Startup Algorithmia" (*GeekWire*, July 2021) https://www.geekwire.com/2021/datarobot-lands-300m-acquires-seattle-machine-learning-startup-algorithmia/ accessed 14 September 2023.

21: Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl, "Machine Learning Operations (MLOps): Overview, Definition, and Architecture" (2023) 11 IEEE Access 31866 DOI: 10.1109/ACCESS.2023.3262138 , 31866.

HUGGING FACE

Hugging Face — founded in 2016 as a computational linguistics startup — managed to break through in the early 2020s as *the* machine learning development platform most widely used by researchers both in and outside of industry. After an initial foray into the development of natural language processing-fueled chatbots, it pivoted toward creating a product commonly described as "the GitHub of machine learning."[22] Oriented initially toward the research community, Hugging Face created a free platform through which third parties could access and/or share datasets, software libraries and pre-trained models.[23] It has also managed to fill a niche as part of the MLOps pipeline, integrating its repositories with ML-deployment infrastructures like Amazon SageMaker. This was strongly facilitated by the release of its open-source 'Transformers' library,[24] used to achieve interoperability across leading machine learning development frameworks like PyTorch (released by Facebook AI Research) and TensorFlow (developed by Google). Hugging Face is funded primarily by a syndicate of the largest AI firms, rather than by more standard venture capital investments, and is a significant player, valued in its last funding round at approximately $4.5 billion.[25]

In a book-length trade publication providing an in-depth look at the ways that organizations and individuals can use their platform, Hugging Face researchers describe what they see as the core added value of their 'model hub':

> In the early days, pretrained models were just posted anywhere, so it wasn't easy to find what you needed. Murphy's law [anything that can go wrong, will go wrong] guaranteed that PyTorch users would only find TensorFlow models, and vice versa. And when you did find a model, figuring out how to fine-tune it wasn't always easy. This is where Hugging Face's Transformers library comes in: it's open source, it supports both TensorFlow and PyTorch, and it makes it easy to download a state-of-the-art pretrained model from the Hugging Face Hub, configure it for your task, fine-tune it on your dataset, and evaluate it.[26]

Hugging Face has grown rapidly, hosting over 2.5 million AI model repositories as of early 2026, with purportedly more than 4 million downloads of its Transformers library every month. The most downloaded models involve a mix of professional and non-professional uploads: image classification models fine-tuned by PhD students, LLMs released by major companies and many other specialized models for translation or computer vision. Leading international research organizations, university institutes and industry labs now disseminate

22: Faustine Ngila, "The GitHub of AI Is Named after an Emoji — and Microsoft Has Its Fingers in It Already" (*Quartz*, May 2023) https://qz.com/hugging-face-microsoft-artificial-intelligence-1850490270 accessed 27 September 2023.

23: Emilia David, "Google, Amazon, Nvidia, and Others Put $235 Million into Hugging Face" (*The Verge*, August 2023) https://www.theverge.com/2023/8/24/23844444/google-amazon-nvidia-hugging-face-generative-ai-investment accessed 25 August 2023.

24: Thomas Wolf and others, "HuggingFace's Transformers: State-of-the-art Natural Language Processing" (arXiv July 2020) DOI: 10.48550/arXiv.1910.03771 accessed 27 September 2023.

25: Kyle Wiggers, "Hugging Face Raises $235M from Investors, Including Salesforce and Nvidia" (*TechCrunch*, August 2023) https://techcrunch.com/2023/08/24/hugging-face-raises-235m-from-investors-including-salesforce-and-nvidia/ accessed 20 September 2023.

26: Lewis Tunstall, Leandro von Werra, and Thomas Wolf, *Natural Language Processing with Transformers* (O'Reilly Media, Inc 2022) xii.

their work directly on Hugging Face.[27] This group now also includes prominent Chinese firms showcasing their open-source wares, despite the platform being cut off by the Great Firewall in May 2023.[28]

Hugging Face has genuinely sought to cultivate a reputation as a responsible actor in relation to ethical concerns including fairness, accountability and transparency. Its researchers have conducted some important critical research on bias in image-generation models,[29] and played an important role orchestrating notable open science projects.[30] The firm has also voluntarily integrated some basic yet nevertheless notable transparency features into its platform design, such as the 'model cards' concept proposed by Margaret Mitchell, hired by Hugging Face after she pseudonymously co-authored a paper critiquing language models (and was fired by Google as a result).[31]

TYPES OF MODEL MARKETPLACES

Hugging Face joins a range of platforms that now commonly host user-generated, downloadable AI models.[32] We have previously termed these model marketplaces:[33] platforms which allow models to be downloaded, so that they can be run on local devices or on any cloud server (see Chapters 2, 3). This distinguishes these kinds of AI hosting platforms from AI-as-a-service providers like OpenAI, which may create interfaces for users to upload or share custom prompts for fine-tuning of their few proprietary products (the 'GPT Store'), but, with a few exceptions, do not allow their models to be freely downloaded.

General-purpose model marketplaces like Hugging Face focus on all models regardless of functionality. Some generic software repositories, such as GitHub, also have model marketplace features and are used commonly for hosting and developing models, albeit in a less specialized manner.[34]

27: Cailean Osborne, Jennifer Ding, and Hannah Rose Kirk, "The AI Community Building the Future? A Quantitative Analysis of Development Activity on Hugging Face Hub" (2024) 7(2) Journal of Computational Social Science 2067 DOI: 10.1007/s42001-024-00300-8 .

28: Joanna Chiu, "New Tool Lets You Track Exactly When the Chinese Government Blocks Websites" (*Rest of World*, 18 September 2024) https://restofworld.org/2024/when-china-blocked-ai-sites/ accessed 14 February 2026. Note that Chinese companies have launched domestic clones of Hugging Face, notably ModelScope from Alibaba, with interfaces designed to mimic both Hugging Face and Civitai, a platform discussed later.

29: Sasha Luccioni and others, "Stable Bias: Evaluating Societal Representations in Diffusion Models" (2023) 36 Advances in Neural Information Processing Systems 56338 DOI: 10.48550/arXiv.2303.11408.

30: Such as the publicly funded language model project BLOOM, see Melissa Heikkilä, "Inside a Radical New Project to Democratize AI" (*MIT Technology Review*, July 2022) https://www.technologyreview.com/2022/07/12/1055817/inside-a-radical-new-project-to-democratize-ai/ accessed 30 September 2023.

31: Margaret Mitchell and others, "Model Cards for Model Reporting" in *Proceedings of the 2nd ACM Conference on Fairness, Accountability and Transparency* (FAT* '19, ACM 2019) DOI: 10/gftgjg ; for the paper that preceded Mitchell's firing from Google, see Emily M Bender and others, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability and Transparency in Computing Systems (FAccT 2021)* (ACM Conference on Fairness, Accountability and Transparency in Computing Systems (FAccT 2021), 2021) DOI: 10/gh677h.

32: Sometimes these are called *open-weight* models, reflecting the fact that their license may not be permissive enough for them to be classified as *open source*.

33: Robert Gorwa and Michael Veale, "Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries" (2024) 16(2) Law, Innovation and Technology 341 DOI: 10/nq2g.

34: These platforms could theoretically incorporate more features oriented specifically toward hosting and sharing models in the future if they see it in their business interest to do so. GitHub has worked with organizations including Hugging Face on policy initiatives promoting open-source AI in the past. See Peter Cihon, "How to Get AI Regulation Right for Open Source" (*The GitHub Blog*, 26 July 2023) https://github.blog/2023-07-26-how-to-get-ai-regulation-right-for-open-source/ accessed

The proliferation of generative models that are optimized to output images or video rather than text has been accompanied by the growing popularity of *image-generation model marketplaces*. Examples of these include Civitai, PixAI, Civision (ModelScope) and Mage.Space. These services provide a platform where users can upload and download fine-tuned versions of a narrower range of models, like Stable Diffusion XL or Black Forest Labs' FLUX, as well as LoRA plugins for them (see Chapter 3). Image-generation model marketplaces are often targeted toward hobbyists and amateurs more than enterprise customers, and bundled with social features such as blogs, instructional forums, bonus points and community challenges to make them attractive and compelling. Users often run these models on-device, using the tools described in Chapter 3.

Finally, *app stores* are worth mentioning. We reviewed the infrastructure that gives power to app stores in Chapter 3, and whether that infrastructure might be conducive to interventions that limit the ability of local devices to engage in malicious activity. Here, we focus on the decision processes that app stores have had to develop when wielding this type of power. Rather than being model marketplaces themselves — one cannot simply upload the raw weights and structure of a model to the Apple App Store for it to be downloaded by the public — there are a growing number of mobile apps that bundle one or more specific models with a user interface to use them, or combine a querying interface with a model marketplace where users can browse and choose models. Notable examples include general-purpose image-editing or creation apps such as *Draw Things*.

We will discuss these three types of AI hosting platforms in this chapter. But let's turn first to an example demonstrating the general challenges of deciding how to govern a model marketplace. Enter the self-proclaimed 'worst LLM ever.'

'THE WORST LLM EVER'

In the summer of 2016, a group of academic computer scientists were regularly scraping data from 4chan, a bulletin-board-style Internet forum regularly mired in controversy. After receiving clearance from an ethics committee at University College London, the researchers systematically collected posts made on 4chan's main politics channel, /pol/, quantitatively characterizing the particular brand of "xenophobia, social conservatism, racism, and, generally speaking, hate" exhibited by users of the notorious platform, which was in the public spotlight in relation to the newly ascendant and online-savvy transnational far-right.[35] They noted that the n-word was used on average more than 120 times per hour on /pol/ alone, and looked to establish a relationship between YouTube links posted by 4channers and subsequent hateful comments on those videos.

In 2020, three of the authors returned to the same social media-focused computer science conference to present a new paper. Their 4chan-scraping infrastructure had been kept up and running since their initial effort, and they had decided to publicly release a dataset with almost

17 November 2023.

35: Gabriel Hine and others, "Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web" (2017) 11(1) Proceedings of the International AAAI Conference on Web and Social Media 92 DOI: 10.1609/icwsm.v11i1.14893.

everything posted on /pol/ from June 2016 to November 2019: more than 3 million threads, and 134 million posts, in a structured and machine-readable *.json* format.[36] This new paper did not mention ethics approval (in general or for the ungated release of the data in particular). Claiming to be following best practices for open data reuse, the dataset was published under a Creative Commons Attribution 4.0 license on Zenodo, a public repository for EU-funded research outputs.

One of the more than 36,000 ensuing downloads of this dataset from the Zenodo archive would land on the hard drive of a computer belonging to Yannic Kilcher, a machine learning PhD student at ETH Zurich and a popular YouTuber with more than a hundred thousand subscribers. In an instructional video that has since amassed over a million views, Kilcher begins by noting that he has "some time and resources, but not nearly enough to train a language model from scratch."[37] Instead, he went to Hugging Face to download GPT-J, a surprisingly powerful open-source LLM created as "a fun hobby project" by a group of bored machine learning researchers hanging out on Discord during the Covid-19 lockdown.[38] He then fine-tuned GPT-J over a two-week period to adapt the model's response style and inject it with additional information from the 4chan dataset.

The result was a powerful large language model that had its statistical weights altered after ingesting an enormous quantity of unbelievably toxic text: a preponderance of slurs and misogyny from a deeply ambivalent, antagonistic space where humor, trolling and racist and sexually violent motifs resonate loudly.[39] In the YouTube video announcing his experiment, Kilcher discusses the end result, which he dubs 'GPT-4chan': "I tested the model... and it was really good... in a terrible sense."[40] Not just a 4channer but also an ML researcher at heart, Kilcher put the model through its paces, assessing its performance on various benchmarks commonly used to compare LLMs. To his delight, it outperformed both OpenAI's GPT-3 and Eleuther's default GPT-J on some benchmarks, in particular the 'TruthfulQA' eval, which had been designed by researchers at Oxford and OpenAI to feature "questions that some humans would answer falsely due to a false belief or misconception."[41] Kilcher's stunt thus also involved an effort to critique the hollowness of dominant ML benchmarking approaches — after all, how could an overtly toxic LLM packed to the brim with a large additional dose of what Kilcher called "very vile opinions" really perform 'better' in some tasks than the current industry standard?[42]

On YouTube, the video description also featured a link to a website created by Kilcher, which allowed users to interact with the model through a Web application. The code he used

36: Antonis Papasavva and others, "Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board" (2020) 14 Proceedings of the International AAAI Conference on Web and Social Media 885 DOI: 10.1609/icwsm.v14i1.7354.

37: Yannic Kilcher, "GPT-4chan: This Is the Worst AI Ever" (*YouTube*, 3 June 2022) https://www.youtube.com/watch?v=efPrtcLdcdM accessed 14 February 2026, 5min 52.

38: Craig S Smith, "EleutherAI: When OpenAI Isn't Open Enough" (*IEEE Spectrum*, 21 March 2022) https://spectrum.ieee.org/eleutherai-openai-not-open-enough accessed 27 January 2026, n.p.

39: Whitney Phillips and Ryan M Milner, *The Ambivalent Internet: Mischief, Oddity, and Antagonism Online* (Polity 2018) 37.

40: Kilcher (n 37) 6min 41.

41: Stephanie Lin, Jacob Hilton, and Owain Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods" in Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (ACL 2022, Association for Computational Linguistics May 2022) DOI: 10.18653/v1/2022.acl-long.229 1.

42: Kilcher (n 37) 0min 52.

to process the dataset for fine-tuning was put on GitHub, and the model itself was shared on Hugging Face. With apparent pleasure, Kilcher described a 'prank' he had run, using the model to power a chatbot that he then let loose on /pol/. This hate speech generator would post anonymously more than 30,000 times on 4chan before Kilcher shut it off.

Users on Hugging Face quickly began to complain that the model would respond to virtually any query with racial slurs, and respond to simple prompts with rabbit-hole conspiracies about wealthy Jewish families causing climate change.[43] But Hugging Face had no structured trust and safety team at the time, and had only generic content policies that did not deal with the distinct challenges posed by AI models, leaving it ill-prepared for complex cases and the trade-offs that these types of uploads imply.[44] There were a few days of back-and-forth discussion on the GPT-4chan repository's public talk pages. Should only the interactive 'widgets' be blocked? Should only logged-in users be allowed to access the model? Should there be a disclaimer attached to it? Following an extraordinary direct intervention by CEO Clément Delangue in the public chat, and considerable internal debate within the firm, the model was taken down.[45]

MODERATING MODELS

Hugging Face struggled to work out what to do with the 'worst LLM ever.' It appeared a little taken aback by the sheer complexity of decision-making in this area, but the underlying reasons are understandable. While the topic has yet to be extensively tackled in academic work, making decisions about which models to allow on a hosting platform brings challenges and trade-offs that can be significantly more difficult than classic content moderation challenges.

The company's content policy journey after GPT-4chan-gate is indicative of the struggles to develop workable software moderation rules. First, Hugging Face attempted to distinguish between 'technical' content, such as models and datasets, and 'human' content, such as comments or discussions. Human content had clear, classic content moderation guidelines, while technical content was instead subject to public discussion, but with no clearly articulated principles or policies against which these artifacts would be judged.[46] This policy then changed to remove the distinction, and was instead replaced by a list of characteristics that could lead to all types of user-uploaded content being 'restricted' (such as to logged-in members) as well as a more process-based set of considerations that might lead to content being 'moderated' (i.e. taken down).[47] In this new policy, Hugging Face staff stated they would now pay attention to the "origin of the ML artifact, how the ML artifact is handled by its developers, and how the

43: Gorwa and Veale (n 33).

44: Hugging Face, "Terms Of Service" (June 2021) https://web.archive.org/web/20210622075735/https://huggingface.co/terms-of-service accessed 17 September 2023, (stating: Your Content must not be misleading or unlawful, and must not violate any of these Terms, applicable law and regulation, or infringe or misappropriate any rights of any person or entity. We may remove your Content at any time, at our sole discretion, if we have a concern about your Content.)

45: Gorwa and Veale (n 33).

46: Hugging Face, "Content Policy" (August 2022) https://web.archive.org/web/20221130213223/https://huggingface.co/content-guidelines accessed 16 September 2023.

47: Hugging Face, "Content Policy" (July 2023) https://web.archive.org/web/20230717150419/https://huggingface.co/content-guidelines accessed 16 September 2023.

ML artifact has been used." Yet again, this read more like a mélange of issues and concepts than a set of predictable rules.[48]

MODEL MODERATION LOGICS: INTENDED, REALIZED AND POTENTIAL USE

Part of the challenge is that there are at least three logics that could motivate governance by platforms that host AI models or other forms of software. And these logics apply whether an AI system is connected to a use that is seen as undesirable by certain stakeholders, or whether it is connected to a use that, like most of what we cover in this book, is pretty universally considered beyond the pale.

First, a platform might, when moderating a model, implement a rule designed to consider the *intention* of the uploader. This is fairly simple from a technical point of view, as it means that the platform doesn't need to examine the complex software itself. Ideally, the uploader would honestly tell the platform what their intention is with the upload, describing clearly the tool they have shared in the model description or model card. Intent requirements are common in criminal law and familiar components of intermediary liability frameworks. However, it might be trickier for a platform to deal with scenarios where the uploader's true intention is expressed elsewhere rather than on the platform, such as posts or ads on other services.

The other two logics are trickier to operationalize. A platform might consider the *realized* use of a model. This is an empirical question, and one that typically kicks in after something goes wrong or a tool is misused, including by people other than the uploader. Accurately assessing realized use is often empirically difficult — proving that a model or other piece of software was even used in a certain incident can be a daunting task. Yet this policy at least gives clear leeway for a platform to react to public scandal. If reporting shows that an AI system was used in a certain way, that could motivate action from a platform. This notion resembles the knowledge requirements that are also often a precondition for criminal offenses.

A platform might finally consider the *potential* use of models. This is analytically difficult, because it is hard to reason about the potential misuse scenarios facing general-purpose software: that list would be inherently somewhat speculative and potentially never-ending. It is also expansive, given the many use cases for different AI models with assorted capabilities; this creates issues concerning how proportionate it is to block a model which has legitimate purposes given scant evidence of actual misuse. This logic has features in common with the *precautionary principle* in technology regulation, an approach that seeks evidence of safety before permitting a potentially risky new product on the market, as well as *strict liability* rules that are common for risky products such as airplanes, where intent or knowledge are ignored when assigning responsibility for accidents.

Various elements of these rationales were acknowledged in Hugging Face's policies. On the one hand, aspects of their rulebook seemed to require intent-based classifications, such as content "designed" or "created for" certain ends. Other parts of the Hugging Face policy were based on realized consequences, such as content that "harms others" or is "used [..] for" other ends. Many categories used the term "promotes," which seemed to encompass both observations of the purpose of the model and considerations of its use in practice. The policies

48: Hugging Face, "Content Policy" (n 47).

give the firm broad discretion to moderate, and provide a helpful steer in terms of expected process. They provide little, however, that Hugging Face could point to in order to justify and defend a specific decision, or to help uploaders anticipate what, exactly, the platform permits. In the GPT-4chan case, the uploader stated that the goal of the model involved alignment or hate-speech classification research, disclaiming the intended use logic. One may consider the deployment of the model on 4chan itself an example of misuse, but this point is evidently debatable. The model more clearly could be said to demonstrate potential risks for misuse, but the company did not articulate a particular threshold or test for making a takedown decision based on that criterion.

FORCED HANDS

For hosting platforms, private rulemaking based on these logics might stem from obligations on hosting intermediaries put in place by law, such as obligations under the UK's Online Safety Act or the EU's Digital Services Act. They might stem from sectoral laws creating liability for certain types of impermissible AI systems. They might also be privately created and enforced as a response to reputational or economic pressures, the power of private infrastructural actors such as app stores or payment processors, the shadow of future legislation or laws which oblige good self-governance.[49] As a result of these political dynamics, platforms like Hugging Face will sometimes be able to write their own rules, but at other times will have their hands forced into a certain kind of action.

Some existing legal obligations arguably already require a platform like Hugging Face to adopt the trickiest approaches to software moderation. In particular, they could require platforms to develop rules around *potential* malicious uses, which will be technically complex, expensive and laborious to enforce. For example, the UK broadly criminalizes the possession or even viewing of information 'useful' to a person preparing a terrorist act, and does not require proof of intent in relation to this information.[50] UK law also regulates files capable of transformation into indecent images that appear to be of children — such as models when given a prompt — in the same highly criminalized manner as conventional child sexual abuse imagery.[51] European data protection law considers machine learning models capable of regurgitating personal data as personal datasets, and may require platforms to moderate for this, even proactively.[52]

49: See generally Ian Brown and Christopher T Marsden, *Regulating Code: Good Governance and Better Regulation in the Information Age* (MIT Press 2013); Robert Gorwa, *The Politics of Platform Regulation: How Governments Shape Online Content Moderation* (Oxford University Press 2024).

50: Terrorism Act 2000 s 58(1)(a). The House of Lords appears to state that this information should be 'designed to provide practical assistance' to a potential terrorist rather than just something that could be just as useful to a terrorist as to another person. See *R v G* [2009] UKHL 13, 43. Yet this does not provide much guidance on technologies that are both as 'general-purpose' and as potentially highly specialized as LLMs are. See further, (published before the UKHL judgment), Jacqueline Hodgson and Victor Tadros, "How to Make a Terrorist Out of Nothing" (2009) 72(6) The Modern Law Review 984 DOI: 10.1111/j.1468-2230.2009.00777.x.

51: Protection of Children Act 1978 s 7(9) (England and Wales); Civic Government (Scotland) Act 1982 s 52 (Scotland); The Protection of Children (Northern Ireland) Order 1978 art 2 (N. Ireland).

52: Case C-492/23 *Russmedia* ECLI:EU:C:2025:935; Michael Veale, Reuben Binns, and Lilian Edwards, "Algorithms That Remember: Model Inversion Attacks and Data Protection Law" (2018) 376 Phil. Trans. R. Soc. A 20180083 DOI: 10/gfc63m ; European Data Protection Board, "Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models" (*EDPB*, 17 December 2024) https://www.edpb.europa.eu/system/files/2024-

At other times, Hugging Face has, quite astoundingly, taken steps to allow *any* actor to give it new rules to enforce. This relates to the platform's support for *licenses* as a form of machine learning governance.

Software licenses are a form of governance that emerged in the 1960s and gained traction as software fell more firmly under copyright protection in the 1980s.[53] Licenses seek to make the reuse of software as intellectual property conditional on the adherence to certain conditions, and have long been used to govern downstream use. Open-source advocates developed notorious 'copyleft' licenses that sought to permit others to build upon or modify software without cost, but only if they did not later release the results under proprietary/for-profit licenses of their own.[54] These were designed to have 'viral' effect, and have been influential on the shape of the development of modern computing. Other licenses have included distribution-related conditions, such as those seeking to prevent military uses of their software,[55] or broad conditions such as the JSON license's statement that '[t]he Software shall be used for Good, not Evil.'[56] There are even satirical licenses, such as the 'Anyone But Richard M Stallman' (ABRMS) license, which provides that individuals can do whatever they want with the software, unless they are Stallman, the developer of the original copyleft 'GNU General Public License.'[57]

Contemporary AI systems are being released under more intricate and atypical license agreements than ever before, specifying in detail how these systems and those based upon them should or should not be ethically used.[58] Hugging Face has been especially vocal in its support of the OpenRAIL family of licenses, seeing them as a promising method of transmitting and enforcing norms in the ML community.[59] The OpenRAIL license, which can be applied to several points in the AI development cycle, permits model reuse, but revokes that right if the licensee or any third party under their control uses the tool for certain purposes, such as to infer certain categories of sensitive data about an individual, predict health characteristics for the purposes of insurance pricing, attempt to predict criminality or synthesize undeclared realistic representations of people or events.[60] Hugging Face has built specific infrastructures into its model repositories, such as 'gated' models kept behind 'clickwrap' contracts, and mecha-

12/edpb_opinion_202428_ai-models_en.pdf accessed 18 December 2024.

53: WS Humphrey, "Software Unbundling: A Personal Perspective" (2002) 24(1) IEEE Annals of the History of Computing 59 DOI: 10.1109/85.988582 ; Amy Thomas, "The First Software Licensing Agreement and Its Relationship with Copyright Law" (*CREATe*, October 2019) https://www.create.ac.uk/blog/2018/11/14/the-first-software-licensing-agreement-and-its-relationship-with-copyright-law/ accessed 6 October 2023.

54: Richard M Stallman, *Free Software, Free Society* (Free Software Foundation 2015).

55: Steve Dierker and Volker Roth, "Can Software Licenses Contribute to Cyberarms Control?" in *Proceedings of the New Security Paradigms Workshop* (NSPW '18, ACM August 2018) DOI: 10.1145/3285002.3285009 .

56: JSON, "The JSON License" (2002) https://www.json.org/license.html accessed 7 October 2023.

57: Landon Dyer, "Another Assembler" (*Dadhacker*, February 2014) https://web.archive.org/web/20140207084017/http://www.dadhacker.com/blog/?p=2106 accessed 7 October 2023.

58: Danish Contractor and others, "Behavioral Use Licensing for Responsible AI" in *2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22', ACM 2022) DOI: 10.1145/3531146.3533143 ; Michael Veale, Kira Matus, and Robert Gorwa, "AI and Global Governance: Modalities, Rationales, Tensions" (2023) 19(1) Annual Review of Law and Social Science 255 DOI: 10.1146/annurev-lawsocsci-020223-040749 .

59: Giada Pistilli and others, "Stronger Together: On the Articulation of Ethical Charters, Legal Tools, and Technical Documentation in ML" in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '23, ACM 2023) DOI: 10.1145/3593013.3594002 accessed 14 June 2023.

60: "Responsible Artificial Intelligence Source Code License" (*Responsible AI Licenses (RAIL)*, November 2022) https://www.licenses.ai/source-code-license accessed 9 November 2023.

nisms for personal data collection, such as the emails of users, in its effort to support uploaders in making their licenses enforceable (Figure 5.1).[61]

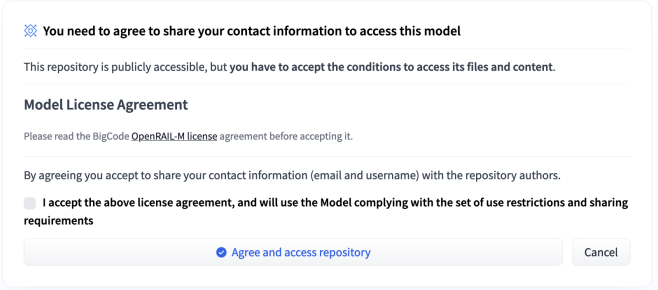

Figure 5.1: Dialog box from the BigCode *starcoder* model, asking users to explicitly accept license terms and forward contact details from users' accounts.

Hugging Face has, at times, integrated violation of licensing into its content moderation decisions. On the one hand, this is designed to allow a more nuanced approach to trade-offs in controversial areas. It is, however, risky to allow third parties to write rules that a platform will need to enforce in already difficult-to-reason-about areas. Hugging Face has previously found itself in hot water concerning the enforcement of upstream licenses preventing 'political' use, for example. The authors of this book found Xi Jinping voice generators that Hugging Face took down from its own platform on the basis of a license violation.[62] However, when journalists followed up as to why these were taken down, while similar Biden voice generators had not been, the platform could not give a principled answer.[63] Licenses are already fairly ineffective means of governance for anyone other than a large corporate actor, because they legally require the license issuer or someone acting on their behalf to do the enforcement themselves, assuming such licensing terms are even enforceable at all.[64] Research has indicated that in the machine learning space, they are often ignored entirely and regularly breached, even by large actors.[65] Enforcement challenges exist even where the simple question is 'is someone using my code?' When the question is transformed into 'is someone *misusing* my code?' the costs of investigation spiral and the feasibility of enforcement wanes further.

ON NOTICE

These rules would not pose a practical issue if they sat there dusty and unused. That, however, is unlikely. Under European law, Hugging Face is obliged to diligently apply terms of service, even though as it stands, the language and structure of the law makes it impossible

61: Gorwa and Veale (n 33).
62: ibid.
63: Joseph Cox, "Hugging Face Removes Singing AI Models of Xi Jinping But Not of Biden" (*404 Media*, 21 November 2023) https://www.404media.co/hugging-face-removes-singing-ai-models-of-xi-but-not-of-biden/ accessed 4 April 2024.
64: Andrés Guadamuz, "Viral Contracts or Unenforceable Documents? Contractual Validity of Copyleft Licenses" (2004) 26(8) European Intellectual Property Review 331.
65: James Jewitt and others, "Permissive-Washing in the Open AI Supply Chain: A Large-Scale Audit of License Integrity" (*arXiv*, 9 February 2026) DOI: 10.48550/arXiv.2602.08816.

to work out what that diligent application would really look like.[66] As a hosting service, it is also entangled in a 'notice-and-takedown' regime, whereby it benefits from legal immunity for liability that such uploaded models might usually accrue, until it is clearly notified of this illegality in a way that, without the need for detailed legal examination, would give rise to 'actual knowledge and awareness' in a 'diligent economic operator.'[67] Where it chooses to take down a model or dataset, or restrict its visibility in any way, it has to disclose why.[68]

However, when it comes to generative AI models, this regime, long standard in the world of search engines and social networks, has serious problems. What information can a notifier provide about the purported illegality of a *model* that would not require a further "detailed legal or factual examination"? Something short and sweet about a model's capabilities — 'This model generates child sexual abuse material when queried' — provided by a flagging user or third-party actor would not be enough. Otherwise platforms would expect a deluge of such notifications, justified or not, leaving intermediaries uneasy about whether or not to leave potentially totally legitimate models up. Of course, an intermediary could test the claims they received, but how would they test such a simple statement? These providers typically have no visibility into how models are used (unless they also operate as cloud providers). Out-of-the-box safety benchmarks are narrow and cannot always be automated, and the capabilities of models can be obfuscated, such as through LoRAs with hidden trigger keywords (see Chapter 3). Scientifically, we lack the ability to peer into a model and reason about how it works, with little successful progress being made in that direction.[69]

Perhaps it would therefore help to have an example of the query that was causing the problem. After the hosting provider got the model to work (which could be a tall order in itself given the infrastructural demands of some of the largest downloadable models), could its staff conduct tests to see if the notifier's claim was correct? Possibly. In some policy areas, such as child sexual abuse and terrorism, even running such a test to generate this content could be a criminal offense in and of itself.[70] The notifier who discovered this kind of content could also potentially be held liable for trying to generate it. And nonetheless, by the time this has all occurred, we would have quite clearly moved into the world of exemptions from "detailed [..] factual examination" that the EU and many other jurisdictions have converged on via statute or case law.[71]

Perhaps in that case, a proper, evidenced, scientifically grounded report should be the basis

66: Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC [2022] OJ L265/1 (DSA) art 14.

67: DSA, arts 6, 16; Case C-324/09 *L'Oréal SA and Others v eBay International AG and Others* ECLI:EU:C:2011:474.

68: DSA, art 17.

69: Lee Sharkey and others, "Open Problems in Mechanistic Interpretability" (*arXiv*, 27 January 2025) DOI: 10.48550/arXiv.2501.16496.

70: For this reason, some have been arguing for good-faith exemptions from liability for providers trying to audit these systems; see Riana Pfefferkorn, "Grok Is Undressing People Online. Here's How to Fix It" (*The New York Times*, 12 January 2026) https://www.nytimes.com/2026/01/12/opinion/grok-digital-undressing.html accessed 27 February 2026.

71: Most intermediary liability frameworks that provide various intermediary safe harbors have some kinds of knowledge standards — either made explicit or interpreted on a case-by-case judicial basis — to remove intermediary immunity. See *L'Oréal and eBay* (n 67). This traditionally requires the intermediary to be informed that they are hosting illegal content, and thus gain knowledge of it, before they are found to be liable for failing to act. While some jurisdictions have sought to in certain cases force intermediaries to do certain forms of proactive screening (and thus actively gain knowledge of prospectively infringing content), this kind of detailed technical examination of models could not be trivially automated and arguably would go much further than existing — already controversial — filtering requirements.

of a takedown request. This would save the intermediary from carrying out the examination itself, as it would be presented with a body of evidence to rely on. And yet, this kind of regime could also be abused: text today is cheap, and the same models that may be subject to takedowns can themselves be used to generate voluminous, plausible reports. Trusted reports might work in such situations, but the free-for-all that currently exists in the law would not. The platform regulation status quo in the EU does provide for "trusted flaggers" whose notices are given *priority*, but it does not provide for a mechanism where they are given more *credence*.[72] As it stands, even the act of getting a platform to act in this space at all is fraught, even assuming that some consensus can be reached on the conditions under which a dual-use model might be considered illegal.

MALICIOUS MARKETPLACES

Hugging Face is a platform that broadly welcomes content moderation, even if it struggles to make sense of exactly how it should work. The former is not always the case, however. Let's now look at Civitai, an image-generation model marketplace established in Boise, Idaho, in late 2022. It is a platform where people upload and download fine-tuned, open-source image and video generation models as well as plugins for them.

The core novel feature provided by Civitai from the outset was an easy-to-use Web interface, which not only features various models available for download and easy deployment, but also large thumbnails depicting select, compelling images generated by each model. Civitai allows users to either download models and query them on their own devices, or purchase a social, gamified token called 'Buzz' that can be used to easily query and fine-tune the models using cloud compute: no coding required.

All of this was combined with familiar design elements from social media — interfaces through which logged-in users can rate, like, comment — as well as novel 'remix' features, allowing users to quickly edit all images generated in the Civitai cloud, immediately loading the model and prompts used to generate them for further manipulation. Blogs, tutorials and competitions also play a prominent role on the site: whereas Hugging Face bridges enterprise software and hobbyists, Civitai positions itself instead as a hands-on technical platform for amateurs with the visual allure of Instagram, the social editing features of TikTok, the camaraderie of Discord and easy-to-use generation interfaces for the technically uninitiated to boot.

Civitai took this model and obtained some modest (by large startup standards) investment in 2023, with a $5 million funding round led by the venture capital firm founded by Netscape Navigator tycoon and provocateur Marc Andreessen.[73] The platform has continued to grow since then, with researchers showing that more than 40 million images were available on the site in early 2025, along with nearly 1,000 standalone models and well over 200,000 plugins or

72: DSA, art 22.

73: Pitchbook, "Civitai Company Profile: Valuation, Funding & Investors" (November 2023) https://pitchbook.com/profiles/company/530262-28 accessed 30 October 2023; Emanuel Maiberg, "Andreessen Horowitz Invests in Civitai, Which Profits From Nonconsensual AI Porn" (*404 Media*, 14 November 2023) https://www.404media.co/andreessen-horowitz-invests-in-civitai-key-platform-for-deepfake-porn/ accessed 16 November 2023.

'adaptors,' such as LoRAs, that could be combined with those standalone generative models.[74]

NONE OF US ARE LAWYERS

At the outset, Civitai appeared to feature virtually no moderation. Conversations with users on Reddit suggested that the platform was a real bootstrapped operation, run by developers without much (or any) policy counsel. In January 2023, posts made by users on the Civitai sub-Reddit recommended that the company publish a privacy policy and a DMCA copyright takedown request form. "Solid idea," the official Civitai account replied.[75] When urged to interrogate the possible legal liabilities and responsibilities that they had for content created and shared via their services, Civitai staff shrugged, noting that "none of us are lawyers," and that despite the distinct challenges of model moderation, their initial terms of service were "copied from standard hosting TOS [terms of service]."[76]

For an ordinary platform where users shared their art, this may not have been such a serious issue. But peering beneath the shiny surface of the platform's front page showed that Civitai's use was overwhelmingly sexual in nature, with research indicating that close to three-quarters of the site's content was classified as not safe for work by the platform's own users.[77] On the large Reddit r/StableDiffusion forum, complaints were appearing about Civitai's 'questionable models' resembling 'underage teenage girls,' and a userbase 'overrun by horny teenage boys.'[78]

In response, the platform had in 2023 half-heartedly put in place a policy prohibiting "[p]ortraying real people in any mature or suggestive context" or depicting real people (such as celebrities) in anything other than "conservative high school dress code."[79] Although this may have seemed promising, the clear practice emerging among the user base was to combine different resources to get problematic or illegal results. One could download plugins making models good at accurately representing a certain real-world individual, ranging from Hollywood stars to minor influencers or one's ex-partner, and chain them together with other plugins intended to make a model good at accurately depicting explicit sexual poses, nudity, certain kinds of body types or particular fetish scenes. Through this technique, the uploaders — and Civitai — could disclaim responsibility for harms arising from users combining together distinct components sourced from the platform.

We have little visibility into what users do with Civitai-sourced models on their own devices, other than what they share via the service later. But we have gotten glimpses of what they do on Civitai's click-to-generate online infrastructure. In 2023, machine learning infrastructure provider and Civitai supplier OctoML conducted an internal investigation and

74: Laura Wagner and Eva Cetinic, "Perpetuating Misogyny with Generative AI: How Model Personalization Normalizes Gendered Harm" (*arXiv*, 20 May 2025) DOI: 10.48550/arXiv.2505.04600 4.

75: https://www.reddit.com/r/StableDiffusion/comments/101j73s/comment/j2ov3hz/ (archived at https://perma.cc/2SZK-YZX8).

76: See https://www.reddit.com/r/StableDiffusion/comments/101j73s/comment/j2ouu5k (archived at https://perma.cc/8BTF-KE5S).

77: Wagner and Cetinic (n 74).

78: https://www.reddit.com/r/StableDiffusion/comments/10rtp17/civitai_alternatives/ (archived at https://perma.cc/KVX7-WW5U).

79: Civitai, "Rules: Real People" (November 2023) https://perma.cc/3YYD-PGX3 accessed 5 November 2023.

found that many of the images being requested by Civitai users, generated by OctoML and then piped back into the Civitai Web application, were at best unethical and at worst illegal, involving both non-consensual pornography depicting real people as well as sexual content depicting minors. OctoML found that even models that were presented on the platform as 'safe for work' turned out to be trivially capable of abuse. One popular model, called 'Deliberate,' could be prompted to produce material that OctoML said could be categorized as synthetic child sexual abuse material by simply asking it to produce images with the prompt "girl: boy: 15, hyperdetailed."[80]

OctoML was worried about the reputational risks of its partnership, likely because the company was in the process of being acquired by Nvidia.[81] It thus began deploying some combination of input and output filtering on queries made by Civitai users, before deciding to terminate its contract with Civitai entirely.[82] Scrambling to respond to the breakdown of some of the key functionality on its site, the Civitai team cheerfully reassured users that "We have found a new potential inference partner, NSFW generation could be back as soon as tomorrow!"[83]

By this time, research had estimated that there were somewhere between 20,000 and 35,000 models or plugins designed to impersonate real people on Civitai, together racking up on the order of 15 million collective downloads.[84] While the platform claimed to flag some repositories for manual approval before upload, it only seemed to do so if the uploader decided to tag the model as 'representing a real person' or 'representing a minor.'[85] It seemed unlikely that anyone intentionally creating synthetic abuse material would voluntarily submit their material for such moderation.

Classic technology funders seemed to be wary of Civitai and the murky waters it was operating in, and at the time of writing, the company has not been able to attract more funding. To survive as a company burning through compute and the substantial costs of hosting and distributing large files, Civitai had to monetize its platform another way. It sought to do so through creating a paid subscription system and a token — 'Buzz' — for image-generation.[86] Although this new business model pulled in close to $2 million in the first year, the platform came under pressure from financial intermediaries processing these transactions: get your act

80: Emanuel Maiberg, "A16z Funded AI Platform Generated Images That "Could Be Categorized as Child Pornography," Leaked Documents Show" (*404 Media*, 5 December 2023) https://www.404media.co/a16z-funded-ai-platform-generated-images-that-could-be-categorized-as-child-pornography-leaked-documents-show/ accessed 2 August 2024.

81: Janakiram MSV, "Nvidia Acquires OctoAI To Dominate Enterprise Generative AI Solutions" (*Forbes*, 30 September 2024) https://www.forbes.com/sites/janakirammsv/2024/09/30/nvidia-acquires-octoai-to-dominate-enterprise-generative-ai-solutions/ accessed 2 February 2026.

82: Emanuel Maiberg, "Civitai and OctoML Introduce Radical New Measures to Stop Abuse After 404 Media Investigation" (*404 Media*, 9 December 2023) https://www.404media.co/civitai-and-octoml-introduce-radical-new-measures-to-stop-abuse-after-404-media-investigation/ accessed 6 March 2024.

83: ibid n.p.

84: Yiluo Wei and others, "Exploring the Use of Abusive Generative AI Models on Civitai" in *Proceedings of the 32nd ACM International Conference on Multimedia* (32nd ACM International Conference on Multimedia, ACM 2024) DOI: 10.48550/arXiv.2407.12876; Will Hawkins, Brent Mittelstadt, and Chris Russell, "Deepfakes on Demand: The Rise of Accessible Non-Consensual Deepfake Image Generators" in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '25, Association for Computing Machinery 23 June 2025) DOI: 10.1145/3715275.3732107.

85: Civitai, "Rules: Real People" (n 79); Civitai, "Rules: Minors" (November 2023) https://perma.cc/CST3-CPNY accessed 5 November 2023.

86: Civitai, "Memberships, Vault, Link, and Generation updates" (April 2024) https://perma.cc/5WJB-XDM7 accessed 4 May 2024, n.p.

together, or have your ability to collect payments through traditional means seriously curtailed.[87]

This again is nothing new. Payment intermediaries such as Visa, Mastercard or PayPal are significant, regular players in content moderation online.[88] They have historically seen sexual content as risky to their finances (due to high chargeback rates associated with online pornography) and to their reputations (due to lobbying by conservative interests).[89] In early 2025, Civitai had its service withdrawn by an unnamed payment processor, almost certainly spooked by non-consensual imagery or CSAM risks.[90] In response, citing also emerging legal uncertainty due to the EU's AI Act and the TAKE IT DOWN Act in the U.S., the platform did what it had until that point been avoiding: it updated its policies to no longer permit any of the popular "[r]eal-person likeness content."[91] There were limits to its compliance, however. Citing regulatory costs, in the same year, Civitai geoblocked itself from the United Kingdom, refusing to comply with the UK's Online Safety Act.

Civitai then began to actively remove many models and images clearly identifiable as depicting real people, whether those were explicit pornography creation models or not. It also announced that it had begun purchasing custom filtering tools from a trust and safety vendor.[92] The Civitai community began to panic, organizing via Discord channels to re-upload the fine-tuned deepfake models en masse to Hugging Face and other hosting platforms as well.[93] Naturally, these resources could also be downloaded and shared in a less-public way through file-sharing and hosting services; a Reddit user on the r/StableDiffusion forum created a website which hashed model files and permitted users to search for the hashes across various hosting platforms, allowing users to find the models they were looking for "across CivitAI, HuggingFace, TensorArt, SeaArt, and torrents" even if they were removed by some of those services.[94]

The end result is still a deeply muddled approach to governing a hosting platform. The overarching trust and safety dynamic on Civitai has shifted away from complete inaction — the company, under pressure from other stack players that it relies on (especially the cloud layer), as well as external, non-AI-stack actors (financial intermediaries), has changed its policies and acted to curb some of the most egregious forms of on-platform misuse. Our simple walkthroughs in January 2026 show that its generation platform does at least basic filtering

87: Emanuel Maiberg, "'Configuration Issue' Allows Civitai Users to AI Generate Nonconsensual Porn Videos" (*404 Media*, 20 May 2025) https://www.404media.co/configuration-issue-allows-civitai-users-to-ai-generate-nonconsensual-porn-videos/ accessed 3 February 2026.

88: Natasha Tusikov, *Chokepoints: Global Private Regulation on the Internet* (University of California Press 2016) (calling them 'macrointermediaries'); Christoph Busch, "Regulating the Expanding Content Moderation Universe: A European Perspective on Infrastructure Moderation" (2022) 27 UCLA Journal of Law and Technology 32.

89: Cole, *How Sex Changed the Internet and the Internet Changed Sex* (n 1).

90: Emanuel Maiberg, "Civitai, Site Used to Generate AI Porn, Cut Off by Credit Card Processor" (*404 Media*, 20 May 2025) https://www.404media.co/civitai-site-used-to-generate-ai-porn-cut-off-by-credit-card-processor/ accessed 3 February 2026.

91: Civitai, "Policy Update: Removal of Real-Person Likeness Content" (May 2025) https://perma.cc/7EGS-RGB5 accessed 23 May 2025.

92: Maiberg, "'Configuration Issue' Allows Civitai Users to AI Generate Nonconsensual Porn Videos" (n 87).

93: Shakeel Hashim, "Why Is Hugging Face Hosting Tools to Make Deepfake Porn of Teenage Celebrities?" (*Transformer*, 5 December 2024) https://www.transformernews.ai/p/why-ethical-ai-company--hugging-face-hosting-pornography-celebrities accessed 13 February 2026.

94: We have decided not to publicly share this URL here, but the reference is on file with the authors.

of prompts, rejecting celebrity names that are not obfuscated or crafted in a way to bypass filtering. However, its proactive moderation has been haphazard at best — reporting published in 2026 shows that various impersonation models are still available on the platform.[95] Our own cursory investigation showed that a number of real-person models were still online and easy to find: a query for 'singer' may not have yielded a Taylor Swift lookalike model, but it did return a LoRA for a Japanese pop star; a simple query for a popular first name returned a pornography LoRA depicting a well-known American actress. The policy change seems to also have simply prompted creators to deploy basic obfuscation on model cards: many of the realistic models we glanced at in early 2026 featured repeated disclaimers that the model weights had been fine-tuned to depict a fictional character, and that any commonality with a real individual was simply coincidental. Unless a human moderator were to happen to recognize the face associated with one of these models, there is no easy way for the platform to assess whether or not the claim to depict a synthetic persona is indeed true without external expertise or outside information.

Simultaneously, the firm has taken measures that show that its business model is at least partly premised on facilitating forms of activity that are at best ambivalent and at worst illegal. In early 2024, it announced the creation of Civitai Vault and Civitai Link, a desktop app to facilitate the easy offline organization of Civitai resources, as well as a cloud-storage solution designed to ensure that "Even if a creator removes a model, it remains safely stored in your personal vault."[96] It is not clear in Civitai's current terms and community guidelines whether models and resources removed by Civitai for policy violations will also be removed from users' Vaults.

MODELS IN APPS

As we have seen throughout this book, committed malicious actors with the need to create a large quantity of high-quality synthetic media now have a range of options available to them. The most savvy and motivated groups might steal cloud credits for ungated access to the best quality multimodal models (Chapter 2); others, seeking to create large quantities of media at scale, might put together a pipeline for adapting models with data of their choice on their own (powerful) devices or via training interfaces provided by platforms like Civitai. But what about more casual forms of misuse?

The model marketplaces we have discussed so far considerably lower the barrier to entry for lay users, but they do not remove the barrier entirely. In the mid-2020s, the easiest way for someone to create a handful of non-consensual intimate images — something illegal in a growing number of jurisdictions[97] — is to simply download a mobile app.[98]

95: James O'Donnell, "Inside the Marketplace Powering Bespoke AI Deepfakes of Real Women" (*MIT Technology Review*, 30 January 2026) https://www.technologyreview.com/2026/01/30/1131945/inside-the-marketplace-powering-bespoke-ai-deepfakes-of-real-women/ accessed 3 February 2026.

96: Civitai, "Memberships, Vault, Link, and Generation updates" (n 86).

97: Clare McGlynn and Rüya Tuna Toparlak, "The 'New Voyeurism': Criminalizing the Creation of 'Deepfake Porn'" (2025) 52(2) Journal of Law and Society 204 DOI: 10.1111/jols.12527 accessed 13 August 2026, 220.

98: Or, navigate to a Web-based application. Combating these kinds of Web apps (noting that many inference frameworks, such as Hugging Face's Transformers library, have Web versions), is a classic challenge given that fewer gatekeepers exist on the open Web. See for example Cassidy Gibson and others, "Analyzing the AI Nudification Application Ecosystem" in *34th USENIX*

'NUDIFICATION'

From the end of 2023, international news outlets began to steadily cover a number of incidents involving teens in Europe and the United States who had used mobile or Web-based 'nudification' or 'undressing' apps on their classmates.[99] These stories were generally slim on technical detail, but the perpetrators did not appear to be particularly tech-savvy individuals piecing together their own software pipelines from Hugging Face, Civitai or other hosting services. When it comes to 'sexual digital forgeries' made with whatever motives in mind,[100] there is, unfortunately, often an app for that: and it may just be a few clicks away if one knows where to look.

App stores operate with a different governance logic than most other platforms. As we explained in Chapter 3, today's most popular smartphones are 'walled gardens,' using infrastructural strategies to tightly control what kind of compiled software can be run on them, creating the 'app store' as a chokepoint for governance in a way distinct from PCs or the Web. While anyone can upload a model to Hugging Face or Civitai, the two dominant app stores (Apple's App Store and the Google Play Store, default for most Android devices) require developers to submit their codebase for review before their app becomes public. Furthermore, compared to the Web, apps have privileged access to the most capable parts of a device's hardware, such as the GPUs needed to run these models.[101]

Both the App Store and the Google Play Store have extensive guidelines, and strict rules on what kinds of content they permit. Apple in particular has long been accused of being overly prudish, in the mid-2000s deciding not to 'tarnish' its brand by permitting pornographic or overtly sexual iPhone and iPad apps.[102] This is the reason why huge brands like Pornhub or OnlyFans don't have mobile apps, even though global monthly consumer spend on OnlyFans far outstrips monthly consumer spend on OpenAI subscriptions.[103] The manual review of apps means that it is unlikely that *explicitly sexual* AI image models packaged into *explicitly sexual* AI apps make it into the dominant mobile app stores.

Instead, app stores have been facing a distinct flavor of the dual-use problem. Open-ended AI apps may advertise themselves on an app store as generic AI image editors, generators or manipulators, but then have hidden deepfake or CSAM creation capabilities, often advertised

Security Symposium (USENIX Security 25) (2025) https://www.usenix.org/conference/usenixsecurity25/presentation/gibson accessed 3 February 2026; Jack L Goldsmith and Tim Wu, *Who Controls the Internet? Illusions of a Borderless World* (Oxford University Press 2006).

99: Sally Weale, "The Rise of Deepfake Pornography in Schools: 'One Girl Was so Horrified She Vomited'" *The Guardian* (2 December 2025) https://www.theguardian.com/society/ng-interactive/2025/dec/02/the-rise-of-deepfake-pornography-in-schools accessed 5 February 2026.

100: For contrasting perspectives on the harms and political economy of deepfakes, see for instance Robert Chesney and Danielle Keats Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security" (2019) 107 California Law Review 1753 DOI: 10.2139/ssrn.3213954; Brenda Dvoskin, "Speaking Back to Sexual Privacy Invasions" (2024) 99(1) Washington Law Review 59 https://heinonline.org/HOL/Page?handle=hein.journals/washlr99&id=65.

101: This may change if the WebGPU becomes more heavily adopted and used, but still the Web is not well-suited for the acquisition and storage of big models in opaque browser storage.

102: Apple has been very vocal in its opposition to the Digital Markets Act rules requiring the company to permit third-party app stores. The first such app store, the PAL Store, exploited the niche to launch with a porn app. See Katie Collins, "iPhones Get Their First-Ever Porn App, and Apple Is Not Happy" (*CNET*, 4 February 2025) https://www.cnet.com/tech/services-and-software/iphones-get-their-first-ever-porn-app-and-apple-is-not-happy/ accessed 5 February 2026.

103: Dan Frommer, "Consumer Trends 2026" (*Coefficient Capital*) https://www.coefficientcap.com/consumer-trends/consumer-trends-2026/ accessed 20 February 2026, 69.

elsewhere, as part of their core value proposition. And it isn't hard for these app developers to find their audience: networks of app developers have marketed these tools to the public by simply using the powerful advertising functionality built into today's popular search engines and social networks. One group of journalists has consistently looked at Facebook/Meta's Ad Library, a tool created by Facebook in 2018 which offers a searchable index of advertisements that fulfill certain criteria and are running on the platform,[104] and found that Facebook allowed more than 25,000 ads for various nudification and undressing apps in 2025.[105] Similar ads have been run on Instagram,[106] Twitter/X,[107] and Google Search.[108]

For such an app to actually create some kind of plausible nude image of an uploaded person, it needs to bundle at least two different models — an image segmentation model that classifies different pixel groupings into categories (e.g. face, clothes) and then what is often called an 'image inpainting' model that can replace certain categories in the image with new pixel groupings.[109] The former requires off-the-shelf computer vision models for image classification; the latter can be well done with open-weight diffusion models like Stable Diffusion or Flux.[110] Because of the difficulty and expense of training these kinds of models from scratch, nudify apps generally rely on a pipeline of open-weight models, and often bundle them directly into the app and run them locally to avoid incurring cloud costs.

Extensive reporting has documented that most of these apps do not explicitly position themselves as nudification apps in the app stores. Some are 'AI avatar' or 'AI girlfriend' apps promising conversations with an LLM;[111] others are positioned as fun 'face swapping' apps, simple image-generators or maybe even slightly more risqué 'AI kissing' image editors.[112] For an app store, assessing the real capabilities of these apps would require either a threat intelligence-style analysis of the developers that create them — whether their Web properties are linked to malicious actor networks, problematic deepfake site domains, whether they are advertising capabilities that are not permitted by app store guidelines elsewhere — or a structured usability audit, which would involve actually interacting with the submitted app and attempting to probe its functionality.

104: Laura Edelson, Tobias Lauinger, and Damon McCoy, "A Security Analysis of the Facebook Ad Library" in *2020 IEEE Symposium on Security and Privacy (SP)* (2020 IEEE Symposium on Security and Privacy (SP), May 2020) DOI: 10.1109/SP40000.2020.00084 ; Paddy Leerssen and others, "News from the Ad Archive: How Journalists Use the Facebook Ad Library to Hold Online Advertising Accountable" [2021] Information, Communication & Society 1 DOI: 10.1080/1369118X.2021.2009002 .

105: Alexios Mantzarlis, "Meta Is still Running Thousands of Ads for AI Nudifiers" (*Indicator*, 21 January 2026) https://indicator.media/p/meta-is-still-running-thousands-of-ads-for-ai-nudifiers accessed 4 February 2026.

106: Emanuel Maiberg, "Instagram Advertises Nonconsensual AI Nude Apps" (*404 Media*, 22 April 2024) https://www.404media.co/instagram-advertises-nonconsensual-ai-nude-apps/ accessed 4 February 2026.

107: Jason Koebler, "Twitter Now Showing Ads For Nonconsensual 'AI Undress' Apps" (*404 Media*, 15 December 2023) https://www.404media.co/twitter-x-now-showing-ads-for-nonconsensual-ai-undress-apps/ accessed 4 February 2026.

108: Alexios Mantzarlis, "Faked Up #12: Google Is Running Ads for Undresser Services" (*Faked Up*, 7 August 2024) https://fakedup.substack.com/p/google-monetizing-undressers-southport-misinfo accessed 4 February 2026; Emanuel Maiberg, "Google Search Includes Paid Promotion of 'Nudify' Apps" (*404 Media*, 6 August 2024) https://www.404media.co/google-search-includes-paid-promotion-of-nudify-apps/ accessed 4 February 2026.

109: Josephine Passananti, Heather Zheng, and Ben Zhao, "Understanding Non-Consensual 'Undress' Applications" (Bergen, NO, CSCW Digital Intimacy Workshop, 18 October 2025) https://people.cs.uchicago.edu/~josephinep/cscw%20talk.pdf.

110: See for instance the workflow in https://stable-diffusion-art.com/inpainting/.

111: Kolina Koltai and Melissa Zhu, "New Companies Linked to 'Nudify' Apps That Ran Ads on Facebook, Instagram" (*Bellingcat*, 18 June 2025) https://www.bellingcat.com/news/americas/2025/06/18/meta-crushai-crushmate-deepfake-nudify-apps/ accessed 4 February 2026.

112: Alexios Mantzarlis, "Faked Up #34: AI Hugging Videos Are a Prelude to More Deepfake Harassment" (*Faked Up*, 29 January 2025) https://fakedup.substack.com/p/ai-gave-us-jesus-hugs-and-automated accessed 4 February 2026.

CURSORY APP REVIEW

A 2026 investigation by the DC-based non-profit Campaign for Accountability actually did the latter, assessing the image-generation capabilities of dozens of 'AI image editors' available in both app stores. By uploading AI-created stock images to these apps and passing simple queries like 'nude torso' to them, the researchers were able to generate non-consensual imagery with 55 apps publicly available in the Google Play Store and 47 on the Apple App Store (38 of which were accessible in both stores).[113] Their estimates, based on publicly available data, suggest that those apps received more than 70 million downloads and generated more than $117 million in revenue for the app developers (and a typical commission of 30% for the app stores).[114]

How is this possible, given Apple's tight-fisted regime, which we described in Chapter 3 as the "most sophisticated array of architectures of control of any consumer computing system"? While the unusual governance arrangements of app stores certainly give them gatekeeping power — most other platforms relying on third-party content do not wish to impose friction on the content creators that make their platform attractive, and thus have relied on 'take down after' rather than 'manually approve beforehand' moderation logics — they also create capacity problems. In 2015, Apple's App Store was reportedly receiving around 1,000 submissions a day; in 2024, that number was up to about 21,000.[115] Google's total submissions in 2024 were even higher, an average of 52,000 per day — almost 20 million total apps that year, of which 2 million were rejected before going public and a further 3.9 million were later removed for various violations.[116] The question of how these firms are coping with this scale remains largely uninterrogated: no empirical work has yet investigated the organizational contours of how app store moderation works.[117] Court documents and public interviews offer a very general picture: in 2019, Apple's App Review process reportedly took around 13 minutes per app, and reviewers scrambled through hundreds of pages of documents to interpret specific guidelines.[118] If this kind of time crunch seems far too short to do a deep investigation of the workings of complex software, it was actually sector-leading — according to court filings,

113: Michael Clauw, "Tech Transparency Project: 'Nudify' Apps Widely Available in Apple and Google App Stores" (*Campaign for Accountability*, 27 January 2026) https://campaignforaccountability.org/ttp-report-nudify-apps-widely-available-in-apple-and-google-app-stores/.

114: ibid.

115: Calculation based on a total of 7,771,599 apps submitted in the year 2024 per Apple's very brief 2024 Transparency Report. Apple, "App Store Transparency Report" (2024) https://www.apple.com/legal/more-resources/docs/2024-App-Store-Transparency-Report.pdf/.

116: Google, "Google Play Annual Transparency Report (2024)" (31 March 2025) https://transparencyreport.google.com/.

117: The two existing papers on App Store moderation, while helpful, provide general reconstructions based on publicly available documentation and media coverage, rather than new analyses powered by primary data collection (e.g. interviews with moderators or policy employees). Other work has undertaken quantitative description of varieties of apps (and e.g. their privacy policies), without investigating the organizational dynamics of app store moderation directly. See Josh Cowls and Jessica Morley, "App Store Governance: The Implications and Limitations of Duopolistic Dominance" in Jakob Mökander and Marta Ziosi (eds), *The 2021 Yearbook of the Digital Ethics Lab* (Digital Ethics Lab Yearbook, Springer International Publishing 2022) DOI: 10.1007/978-3-031-09846-8_6; Esther Weltevrede and others, "Gatekeepers of the Mobile Ecosystem: Understanding App Store Moderation" (*Social Science Research Network*, Rochester, NY, 30 May 2025) DOI: 10.2139/ssrn.5292303 ; Reuben Binns and others, "Third Party Tracking in the Mobile Ecosystem" in *Proceedings of the 10th ACM Conference on Web Science* (WebSci '18: 10th ACM Conference on Web Science, ACM 15 May 2018) DOI: 10.1145/3201064.3201089 .

118: Philip Shoemaker, "WWDC Preview With Apple's Former App Approval Chief" (*Bloomberg*, 28 May 2019) https://www.bloomberg.com/news/audio/2019-05-28/wwdc-preview-with-apple-s-former-app-approval-chief-podcast accessed 20 October 2025, 8m46s.

Samsung's Galaxy Store reviewers were expected in 2023 to look at 85 apps a day in a 44-hour working week, an average of 6 minutes per app, typically never even loading them and only using metadata, names, tags, descriptions and screenshots to investigate their functionality.[119]

Some very basic additional information on the labor of app governance has recently become available, as the Play Store and the App Store have been classified as 'Very Large Online Platforms' by the European Commission, making them subject to some special transparency reporting requirements under the Digital Services Act.[120] The App Store's public 2024 report notes that it has 606 'human resources dedicated to content moderation' for all their European markets, although it is not clear what proportion of these employees or contractors actually conduct app reviews (in the content moderation sense we're discussing here), since the report also discusses other peripheral moderation functions, such as evaluating the ratings and comments made by users on the apps that they have downloaded.[121] It is also unclear if these 'human resources' work exclusively for the European market or are organized as part of a global network of app moderators (and therefore, it is not evident how many global app reviewers the firm has). Google's 2025 DSA report notes that the firm has 1,110 human moderators reviewing for the European Play Stores.[122]

The app stores have the power, in theory, to make developers wait as long as needed to conduct a thorough evaluation of submitted apps. But in practice, being a global gatekeeper is a huge amount of work, no matter how unbelievably lucrative it may be. It also involves political economy and business model dynamics that can be orthogonal to the safety aspects of app reviewing. There are numerous anecdotal examples of App Review teams sometimes 'missing' violations intentionally, and choosing to turn a blind eye to particularly profitable apps that don't overtly break the rules. Apple's head of fraud, Eric Friedman, once said as much in an internal email: "as much as those [App Review] guys want to help, their paycheck depends on getting apps in the door and keeping developers happy. They are more like the pretty lady who greets you with a lei at the Hawaiian airport than the drug sniffing dog who, well, never mind."[123]

App stores don't want to make their review processes *too* onerous or time consuming for the many developers that they rely upon to create useful software for their mobile ecosystems, as long as those apps don't pose a direct threat to their business model (as the 'apps-within-apps' discussed in Chapter 3 once did). But they do also face a version of the dual-use problem also facing Hugging Face and other hosting services: apps with AI functionality can simultaneously exhibit both innocuous image-editing and manipulation functionalities while also potentially allowing for the creation of CSAM (illegal in numerous jurisdictions) or non-consensual im-

119: *Montres Breguet SA & Ors v Samsung Electronics Co Ltd & Anor* [2023] EWCA Civ 1478, 32–33.

120: DSA, art 42.

121: Apple, "DSA Transparency Report: App Store" (*Apple Legal*, April 2024) https://www.apple.com/legal/dsa/transparency/eu/app-store/2404/ accessed 5 November 2024.

122: See Google, "EU Digital Services Act Biannual VLOSE/VLOP Transparency Report (January-June 2025)" (28 August 2025) https://transparencyreport.google.com/ This figure is inferred, rather than explicitly stated, in the report; the firms break down their moderator numbers by language spoken, rather than total numbers of humans, and tend to have reviewers with baseline skills in English as well as a second or third EU language. For example, Google discloses one Finnish-speaking moderator, but does not clarify if this moderator also evaluates in English (and thus is counted in the 1,110 figure for English) or not.

123: *Epic Games, Inc. v. Apple Inc.*, No. 4:20-cv-05640, (N.D. Cal. May 28, 2021) (Epic Games, Inc.'s Administrative Motion to Seal Portions of Its Proposed Findings of Fact and Conclusions of Law, ECF No. 777), 397.

agery (illegal in a growing number of jurisdictions).

The simplicity of prompts passed to these AI image apps in the Campaign for Accountability's toy audit suggests that app stores are not currently doing any kind of active examination of the AI features integrated into apps. (App stores have also been quite negligent when it comes to sexual impersonation more broadly, with reports that they have allowed developers to pay and advertise banned features, such as 'undressing' or 'nudification,' *in the App Store itself*, so that innocuous-seeming apps will appear as suggested results for those problematic queries. This isn't exactly a subtle gambit to get past the all-seeing gatekeepers!)[124] However, if Google and Apple were to crack down in the future, it is entirely feasible that these same apps could simply obfuscate their problematic features even further, requiring certain prompts or keywords to trigger hidden, potentially illegal functionality (see Chapter 2). Proactive investigation/red-teaming by app stores during the app review process would catch the most blatant apps, but as excellent open-source-intelligence reporting on Web-based nudify apps has shown, their developers are motivated by an extremely lucrative multi-million-dollar market, and will go to great lengths to evade sanction and pose as legitimate businesses (e.g. to process payments).[125]

The only real way out of this problem for app stores — a way they have so far resisted to protect their revenue streams — might be a *prima facie* ban on all AI image and video editing apps. While that could make it more difficult for someone to search in the app store and experiment with various 'face swapping' apps until they find one that will do the more problematic things that they are looking for, it could also feasibly spiral into an endless game of whack-a-mole, as those apps could stop branding themselves as AI apps and instead become simple image-editing or retouching apps. Might app stores need to ban all media manipulation apps too?

It might not just stop with image editing apps: sophisticated actors can and will deploy ways to get around app review and obfuscate what an app does. Perhaps most infamously, Uber deceived Apple about its use of a private API to read the phone's permanent serial number, supposedly to stop fake Uber accounts that were being used to get sign-up bonuses, by audaciously fighting technoregulation with technoregulation: it included a piece of code which looked at the phone's location. If the phone detected it was in Cupertino, the location of Apple's headquarters where the checks are reputed to take place, the offending API use would magically vanish.[126] Savvy malign image app developers could feasibly deploy similar techniques to bundle problematic AI functionality into any kind of app, and then advertise those capabilities off-platform. Suppose a teenager, raised in a misogynistic culture and unable to realize what a terrible idea generating fake nude photos of a classmate is, comes across an ad on social media or in a gaming chatroom nudging them toward the app store. Given the risk of being caught out by parental credit card statements or phone checks, a malicious app disguised as a seemingly innocuous calculator might be all the more appealing.

124: O'Ryan Johnson, "Nudify Apps Get Past Google, Apple App Moderation" (*The Register*, 27 January 2026) https://www.theregister.com/2026/01/27/nudify_app_proliferation_shows_naked/ accessed 4 February 2026.

125: Kolina Koltai, "Behind a Secretive Global Network of Non-Consensual Deepfake Pornography" (*bellingcat*, 23 February 2024) https://www.bellingcat.com/news/2024/02/23/behind-a-secretive-global-network-of-non-consensual-deepfake-pornography/ accessed 4 February 2026.

126: Mike Isaac, "Uber's C.E.O. Plays With Fire" (*The New York Times*, 23 April 2017) https://www.nytimes.com/2017/04/23/technology/travis-kalanick-pushes-uber-and-himself-to-the-precipice.html accessed 30 August 2025.

EMERGING GOVERNANCE LOGICS

Platforms at the hosting layer all face slightly different combinations of the same problems. The first is the dual-use problem that has been a constant theme in this book thus far: AI systems have dynamic, open-ended capabilities, and actually determining the *potential* for misuse of those systems is often not easy for hosting providers with limited visibility into what that system is actually doing when it is deployed. The second is the problem of imperfect information; large platforms with thousands or even millions of users need to sniff out the signals in a huge amount of noise to do effective trust and safety, and their staff may need to do complex, proactive investigations to find both models/actors uploading those models with the potential to be misused as well as evidence of actual misuse. Finally, there is the problem of obfuscation: motivated actors will seek to exploit the dynamics of the previous two problems to hide their intentions and the capabilities of their tools, seeking, once moderation comes, to evade those that would disrupt them.

General-purpose, public hosts like Hugging Face and GitHub need to grapple with a genuine dual-use problem, as they facilitate access to many extremely open-ended, general-purpose models that can have legitimate, creative uses in research and beyond, but also have troubling capabilities. While GPT-4chan may seem like an extreme case compared to the releases from the leading firms that all have models on the site, Hugging Face is (as of early 2026) home to an enormous number of 'uncensored' models with inherently high misuse potential. These are generally uploaded by individuals, rather than research labs, after tinkerers have removed model guardrails through a process of *abliteration*, a process we discussed in Chapter 3. A cursory search in February 2026 reveals that Hugging Face returns 5,297 results for 'abliterated' models.

AI models — and adaptors created to customize those models — are complex mathematical assemblages, inherently difficult to interpret and interrogate. A firm like Hugging Face thus faces the same observability problem described in Chapter 3: if the underlying purpose of a model or plugin is not explicitly disclosed by the uploader, understanding whether a model or plugin was, in fact, trained on illegal material to do illegal things, can be very difficult without broader context. General-purpose hosts might have hundreds or even thousands of such artifacts uploaded per day.

Even the biggest and best-resourced general-purpose hosting platforms have yet to take the ensuing information problem seriously. There appear to be hundreds, if not thousands, of real-person likeness models still easily available on Hugging Face, many of which are advertised as being NSFW models on external sites that link to the Hugging Face repositories, and are not particularly hidden (with filenames like `taylor-swift.safetensors`). As of February 2026, Hugging Face does not have a clear policy against AI artifacts that do real-person impersonation (or even, as a minimum, are advertised as impersonating a single individual).[127]

127: As of Feb 2026, section 4 of Hugging Face's content policy prohibits "Content that violates the privacy of a third party, including but not limited to publishing others' private information, such as a physical or email address, without their explicit permission," and "Content that infringes the intellectual property rights of a third party." They also have further rules on "intentionally deceptive" or "fraudulent" content. Using someone's likeness without permission is notably missing here, although this could perhaps be interpreted by the company to remove a model if they specifically got a takedown request for it. The problem of how Hugging Face would deal with models that are not explicitly impersonation resources and yet still impersonate certain public figures (e.g. most open-weight diffuser models) is also not clear, and perhaps for this reason they do not wish to wade

This suggests that the platform is doing little to no proactive moderation or red-teaming to remove the most blatantly problematic deepfake models. Much like the Apple App Store allowing generically branded AI apps that can be trivially prompted to produce potentially illegal material, this failure does not exactly raise the complex problems of interpretability that we have discussed throughout this book: these platforms could simply look and remove the most obviously violative models/apps and users if they had a meaningful desire or incentive to do so.

Civitai and other similar image-generation-oriented hosts also face some elements of the dual-use problem. Civitai, under pressure from off-stack actors — in particular, financial intermediaries and infrastructure providers — has begun to do more to disrupt obvious forms of on-platform misuse, removing most blatantly policy-violative models and LoRAs, and also implementing some filters and other measures to safeguard the integrated cloud-provided services it offers. It is unclear to what extent it has also rolled out these measures for the 'censorship-resistant' features that it has itself made available to users (e.g. downloadable offline model libraries), and its business model involves walking a tightrope between funders and financial intermediaries on one side and the whims of a community that seems to enjoy what are at best problematic and at worst illegal activities on the other. While previously unpoliced platforms may have once had blatant misuse happening out in the open, the crackdown on overtly problematic AI artifacts by these platforms will likely just drive users to obfuscate their activity — not providing extensive descriptions on their LoRAs, for instance, and obscuring their real trigger words and functionality — or push them off-platform. Empirical research will be needed to gauge the real efficacy of these interventions, but given the role that some of these services have played in the past in facilitating the easy access to sophisticated deepfake models for non-technical users, this added friction is probably a positive development.

App stores are primarily facing a combination of the information and obfuscation problems. They are stressed by the sheer quantity of app review that their central gatekeeping role has now come to require, and in many cases perform cursory app approvals that allow actors to very easily obfuscate the real functionality of the software they have created — functionality that can bring in large, lucrative customer bases through off-platform advertising. Such advertising appears very difficult for even the largest online ad monopolies to deal with. Facebook, which for years has resisted calls to approve new advertisers or potentially risky advertisements manually, was unable to build effective automated systems to stop some particularly pesky nudifier developers from advertising on its platforms. It has since taken direct legal action against them, and deployed some additional measures that we will discuss shortly.[128] The growth of AI-powered 'vibe coding' has further swollen app submission rates, making moderation harder still.[129] The latest reporting suggests that the app stores have remained in

into this territory. In response to previous work by the authors and by 404 Media, they did tell reporters that they *had* a policy against impersonation without consent, but we could not find what they were referring to. See Cox, "Hugging Face Removes Singing AI Models of Xi Jinping But Not of Biden" (n 63). The policy is available at https://huggingface.co/content-policy, archived at https://perma.cc/5LB6-3LQZ.

128: Maxwell Zeff, "Meta Sues AI 'nudify' App Crush AI for Advertising on Its Platforms" (*TechCrunch*, 12 June 2025) https://techcrunch.com/2025/06/12/meta-sues-ai-nudify-app-crush-ai-for-advertising-on-its-platforms/ accessed 12 February 2026.

129: a16z, "Charts of the Week: The Almighty Consumer and AI Capex" (16 January 2026) https://www.a16z.news/p/charts-of-the-week-the-almighty-consumer accessed 20 February 2026.

a reactive stance, taking action when faced with evidence of misuse presented by journalists or researchers, but doing little in terms of comprehensive or proactive governance that would prevent these apps from being easily uploaded to their services.

CODES AND CONDUCT

The pressure on firms at the host layer of the AI stack has been rising since we first began our work on the topic in 2023. In fall 2025, the Australian eSafety Commissioner published a series of binding codes of conduct under the powers delegated to it under Australia's 2021 Online Safety Act. These codes now include a definition of a 'model distribution platform' and a basic set of requirements — primarily, that platforms have contact forms or emails available, and that they "[h]ave systems and processes, including standard operating procedures to enforce their policies when they become aware of non-compliance with the policies."[130]

In response, Hugging Face added a single paragraph to its content policy:

> *Before uploading any content to the Hub, particularly any ML Artifacts, you must take appropriate steps to minimise the risk that the content could be used to generate child sexual exploitation material or pro-terror material.*[131]

This policy change, and the Australian intervention that prompted it, make no mention of any kind of proactive governance efforts — the forms of simple investigation that could get rid of the most blatantly violative LoRAs, models and datasets on the platform. A deeper version of this (beyond the current requirements in Australian law) might actually involve red-teaming and inspecting the actual resources and their functionality, undertaking basic tests on high-risk models, LoRAs and apps. Some European think tanks have, uninvited, begun doing this work externally for Hugging Face, passing simple prompts to leading models deployed publicly for anyone to access on Hugging Face's own paid 'Spaces' infrastructure, showing that these dual-use models trivially undressed researcher-provided stock photos.[132] Supremely wealthy app store operators like Apple could surely afford to bring on this expertise themselves; smaller hosts would benefit from either building proactive analysis capacity in-house, or making use of this capacity periodically if it were to be collectively provided or funded by a cross-industry entity or network.

The question of how to actually bind platforms to take this step, however, is a more complicated one. EU law has a prohibition on member states requiring platforms to generally monitor their services, which heavily limits obligations shaped like proactive duties.[133] Perhaps more importantly, social media law has fallen into a scale trap.[134] Laws such as the UK's

130: eSafety Commissioner, "Schedule 6 Designated Internet Services Online Safety Code, Class 1C and Class 2 Material" (2025) https://perma.cc/YCC9-U93U accessed 2025, 11.

131: Hugging Face, "Content Policy" (January 2026) https://web.archive.org/web/20260131134739/https://huggingface.co/content-policy accessed 12 February 2026, n.p.

132: Paul Bouchaud and Silvia Semenzin, "Unmoderated by Design: How Hugging Face Enables NCII" (*AI Forensics*, 28 July 2026) https://aiforensics.org/work/hugging-face-ncii accessed 13 August 2026.

133: Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce') [2000] OJ L178/1 (e-Commerce Directive).

134: This parallels the way that much-ballyhooed regulatory approaches relating to AI model development have completely

Online Safety Act and the EU's Digital Services Act reasonably make the biggest platforms follow more obligations than the smallest ones. But Hugging Face and GitHub are, by the commonly used measures, small platforms, with relatively few 'monthly active users.' Their primary users are developers, who, despite their best efforts to amplify their productivity via an unquenchable thirst for LLM tokens, nonetheless represent a tiny proportion of the world's population. Yet the upshot can be huge; the content placed on model marketplaces has serious impacts upon software supply chains and increasingly the world more broadly. In other words, AI hosts are an unusual creature largely unanticipated by extant platform law: very important modestly sized platforms.

As a broader trend, hosting providers that currently allow anyone to upload material with little to no friction could also feasibly move toward the app store model, building in approval processes for newly uploaded AI artifacts. Hugging Face should probably do this: it would not have to be as expensive and bureaucratic as the reader might initially imagine. Along the lines of what we discussed in Chapter 3, such processes could involve segmenting users and content in various ways. Major general-purpose hosts could give reputable research labs permission to share resources as they wished, while requiring ordinary individuals to go through a review process or build up certain types of reputational indicators before being granted permission to bypass the ordinary approval channels, or, in a weaker version, to have their uploads made resilient against notice-and-takedown efforts. This kind of approach is not unheard of in the classic user-generated content space: for instance, some social networks, balancing the significant potential of livestream video to facilitate harm with the significant costs of moderating it effectively, have in the past limited those features only to channels that have amassed certain follower thresholds and built up sufficient 'user histories.'[135] The most tightly moderated pornography platforms, like Pornhub, invert the classic 'receive notice and review material before taking it down' model, taking down content *immediately* following any report that it depicts non-consensual material, only reinstating it after a manual review has determined that it was a false flag.[136] App stores, while admittedly facing an absolute torrent of app submissions, could easily segment higher-risk image editing and generating apps into a specialized reviewing process that would require developers to submit documentation about the filters built around any integrated model, provide evidence of red-teaming efforts that show that their models do not respond to the most obvious illicit prompts and subject their submitted apps to periodic direct testing by app store moderators to ensure that these interventions are not simply reverted once an app is approved.

Another feasible governance intervention would be to improve certain forms of infra- and cross-layer coordination in the AI stack, allowing hosting platforms to share resources with other AI intermediaries so that they could more robustly combat the dispersal and diffusion of

misunderstood the relationship between scale and risk, setting requirements based on measures like 'floating-point operations per second' (FLOPS) even though there is little correlation between those measures and the extensive current, already-exhibited societal risks of models.

135: Robert Gorwa and Dhanaraj Thakur, "Real Time Threats: Analysis of Trust and Safety Practices for Child Sexual Exploitation and Abuse Prevention on Livestreaming Platforms" (*Center for Democracy and Technology*, 21 November 2024) https://cdt.org/insights/real-time-threats-analysis-of-trust-and-safety-practices-for-child-sexual-exploitation-and-abuse-csea-prevention-on-livestreaming-platforms/.

136: See the section on 'content removal requests' in Pornhub, "2025 Transparency Report (Second Half)" (*Pornhub Help*, 31 December 2025) https://help.pornhub.com/hc/en-us/articles/50069808248339-2025-Transparency-Report-Second-Half accessed 13 August 2026.

the most blatantly illegal models and resources. Hashes of models are already broadly available to allow users to verify what they download, particularly in the context of malicious code,[137] but these could be used more widely to propagate reputation and moderation decisions. This has its limits — changing the hash of a model without changing its functionality is currently trivial.

Yet making models more legible can be a double-edged sword in an arms race. Non-consensual image production networks seeking to evade host platform moderation have also capitalized on model hashing, building services that use hashes to serve as a model 'search engine,' providing a 'central hub to find information about models, including those previously removed, see all its known online locations (including torrent sites), and get direct links to download it at the source.'[138] Hosting platforms could trivially also share hashes of models confirmed to be illicit if they wished. For instance, if Civitai had given Hugging Face and GitHub the SHA256 hashes of the thousands of adaptor files it removed from its platform in spring 2025 due to its crackdown on non-consensual pornographic impersonation,[139] those platforms could have more effectively removed those same files from their services, or at the very least escalated them for potential moderation and intervention. While users seeking to spread those files could certainly modify them to change the hash values and thwart this kind of matching, this would nonetheless insert friction that would make the spread of these sorts of resources at scale more difficult and time-consuming.[140] Were more robust hash technologies to be developed for the purpose of re-identification rather than verification, this friction could be increased.

Coordination mechanisms could also help hosts deal with the off-platform information problem. There are limits to how much visibility app stores can be expected to have into how app developers are describing their products in other corners of the Web (although looking at open archives of advertisements, such as those operated by Meta or other social networks, isn't actually that difficult). But if those cross-stack actors are conducting their own investigations and also sharing information about their interventions, that can be a helpful hook to grab onto later when looking at, for example, a particular developer or user's history and making a decision about an app or a model.

This kind of sharing is also not entirely unprecedented. In summer 2025, Meta announced a series of measures it had taken against a particularly pesky advertiser — the Hong Kong-based firm Joy Timeline HK Limited — that had created multiple Facebook and Instagram accounts to run thousands of ads for its Web-based nudification app 'Crush AI.'[141] Alongside

137: David Cohen, "Data Scientists Targeted by Malicious Hugging Face ML Models with Silent Backdoor" (*JFrog*, 27 February 2024) https://jfrog.com/blog/data-scientists-targeted-by-malicious-hugging-face-ml-models-with-silent-backdoor/ accessed 12 February 2026.

138: To avoid this quote easily leading people to this service, given that it indexes many malicious models, we have chosen to use the classic Internet research technique of 'bricolage' and rephrase this citation to make it less discoverable. Contact the authors if you need the information for research. For an ethical discussion of this technique, see Annette Markham, "Fabrication as Ethical Practice: Qualitative Inquiry in Ambiguous Internet Contexts" (2012) 15(3) Information, Communication & Society 334.

139: Hashim (n 93).

140: In this particular case, actors might compress large numbers of LoRAs to evade detection, which in and of itself might be a suspicious signal to spur further investigation, but are probably unlikely to actually re-train those LoRAs at the scale of hundreds or thousands of models.

141: Zeff (n 128).

stated changes to its automated ad analysis systems, Meta sued Joy Timeline in Hong Kong court and also began sharing URLs of various Crush AI nudification properties with other participants in the Tech Coalition,[142] a little-studied policy network and tech consortium that helps the industry coordinate its governance efforts against child sexual abuse material (more on this organization in the next chapter). The Tech Coalition's Lantern data sharing infrastructure allows firms to pool not just hashed content — going beyond precursors in other domains, like terrorism[143] — but also user-level information (IP addresses, usernames, email addresses and other 'signals' that could fuel further investigation). Interestingly, Meta's intervention suggests an expansion of this initiative beyond the CSAM context and into slightly more ambiguous, less universally illegal malicious contexts like non-consensual imagery generation. Even if it were to remain more tightly scoped, hosting platforms could surely also participate in this consortium to share hashes of CSAM-producing adaptors and artifacts, and it seems perfectly feasible that this consortium, or something similar, could help hosts direct their scarce moderation resources in more efficient ways.

IF IT IS INDEED BROKE...

These types of interventions can and will be worked around by adversarial actors. They could also lead to some governance capacity that could be weaponized by firms (or, indirectly, by governments or rightsholders putting pressure on those firms) to remove legitimate models, and limit the actions of certain users and developers. Careful institutional design would be needed to curb the potential for coordination mechanisms themselves to have major unintended consequences.

But these kinds of hosting-layer governance interventions could feasibly disrupt a sizable proportion of genAI-powered malice, especially that facilitated with image-generating technologies and predicated on violations of sexual privacy. Hosts could do more to combat overtly problematic models and adaptors; it's less clear what they can and should do to robustly intervene in areas like fraud and romance scamming, where the core functionality desired by the fraudsters can be gained from legitimate open models and doesn't necessarily require specialized, somewhat identifiable adaptors.

The efforts of major platforms to simply deploy licensing and clickwrap contracts to diffuse their own responsibility are ineffectual at best. Because of their scope and operation, OpenRAIL and similar licenses also do little to help platforms navigate the questions of what to do with the hosting of dual-use technologies unless derivative versions of them breaching the license are hosted on the platform too. Even where this is the case, it assumes that the risks come from the fine-tuning rather than the original, base models themselves. Even assuming certain categories of models are obliged to be accompanied by behavioral use licenses, platforms will struggle to find means to oblige the license issuers to enforce those licenses, leaving their enforcement up to the platform itself. In those situations, licenses are just more

142: Meta Public Policy, "Combating Nudify Apps with Lawsuit & New Technology" (*Meta Newsroom*, 12 June 2025) https://about.fb.com/news/2025/06/taking-action-against-nudify-apps/ accessed 12 February 2026.

143: Marguerite Borelli, "Social Media Corporations as Actors of Counter-Terrorism" (2021) 25(11) New Media & Society 2877 DOI: 10.1177/14614448211035121 ; Robert Gorwa, "The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content" (2019) 8(2) Internet Policy Review 1 DOI: 10.14763/2019.2.1407 .

expensive, less consistent and more narrowly applicable forms of content moderation.

In general, hosts seem hesitant to act against AI resources. While the potentially beneficial use cases do cause a conceptual headache for action, there are times that this seems more of an excuse than a reason. In the case of nudification tools in particular, app stores could either ban AI apps or enact higher standards for them, rather than only acting in a piecemeal and reactive manner. They may be worried about dampening AI hype, or fearful of closing off a future income stream. Whatever the case, they have not moved far in this direction. Similarly, Hugging Face has allowed thousands of 'uncensored,' 'abliterated' models to stay on the platform (which are certainly risky, if also feasibly useful in certain contexts), and also seems to tolerate deepfake models and LoRAs that do not overtly brand themselves as malicious.

To conclude this discussion for now, we want to forecast some of the potential longer-term effects on the AI ecosystem posed by a growing role in AI governance by host platforms. The hosts we have discussed are among the largest, most visible players in the generative AI space — arguably the most impactful, but far from the only nodes in a global network of AI hosting providers. If these hosts start to become major governance actors, we are likely to see the migration of malicious users to more difficult-to-reach corners of the Internet — where they will share problematic AI models and artifacts via file-sharing sites, generic cloud storage services, server-to-server transfers or other offline techniques. It's very hard to stop someone from putting a hard drive in the mail.

The most malicious, egregious and illegal toolkits (optimized for creating realistic CSAM of real individuals, for instance) may be removed from public hosts and spread in the Web's shadowy underbelly, moving to smaller, less-known, less compliant platforms, and spaces harder to reach due to their legal jurisdiction or technical characteristics. In 2024, we wrote briefly about *Shoggoth*, a general-purpose model marketplace built on censorship-resistant peer-to-peer technologies, and designed explicitly with cyberlibertarian ideals to foreclose the moderation and regulatory-access-point potential of the platforms we have described in this chapter.[144] Like much of the Internet in the short-lived Web3 heyday of the 2020s, Shoggoth itself was soon consumed by its own libertarianism, descending into a failed cryptocurrency tokenization effort, yet its legacy lives on in many similar projects. Magnet links used to initiate BitTorrent downloads of leaked and uncensored models circulate widely on the Internet. Meta's LLaMA model series had its staggered release foiled in 2023 by the rapid upload of BitTorrent links to download the model on 4chan and similar forums; toolkits now exist to easily mirror Hugging Face models and similar onto BitTorrent en masse, with just a few clicks.[145] Moderating model marketplaces will not completely remove the circulation of malicious models — however defined by the stakeholders seeking to govern them — on the Internet. But such models will certainly become harder to find, and harder to share with others.

144: Gorwa and Veale (n 33).

145: See, for example https://github.com/Nondzu/LlamaTor, as well as the use of IPFS for distributed file-sharing.