Chapter 3

Local: Generation and Regulation On-Device

Michael Veale and Robert Gorwa

Chapter 3 of *Malice, Models and Middlemen: The Full Stack Fight Against AI Misuse*, pages 56–92

Abstract

Consumer-grade hardware has become increasingly adept at running the type of AI systems capable of being seriously misused. In this chapter, we illustrate this through case studies of scam compounds and their associated underground tool marketplaces, and online communities for the generation of synthetic child sexual abuse material. This poses a problem, as regulating devices is more difficult and in many ways more invasive than governing cloud services, given the private nature of activities on these devices and the more limited control and visibility rule-setters have over them. Despite this, building in limits on the software that can run on individuals' computers has a long history, stretching from the Nintendo Entertainment System in the mid-1980s to the iPhone today. We identify and characterize two approaches to regulating activities on devices: software gatekeeping, most familiar from the App Store, and operating system surveillance, which has been at the heart of debates on 'client-side scanning.' We consider how these might apply to the abuse of AI systems, examining the theoretical possibilities and limitations of approaches including notarization of models, restricting access to on-device machine learning hardware and behavioral monitoring. Given that extending these approaches to every device on Earth is unrealistic, we draw on other areas of technology policy where there have been successful efforts at segmenting 'safer' publicly available devices from more specialized open-ended ones. Careful combinations of market design and technical interventions could be deployed to offer a limited — if thorny and potentially dangerous — path forward to disrupting the most harmful forms of locally facilitated AI misuse.

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 3

https://homepages.ucl.ac.uk/~ucqnmve/malicemodelsmiddlemen/chapter/3/

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.

3 LOCAL: GENERATION AND REGULATION ON-DEVICE

Consumer-grade hardware has become increasingly adept at running the type of AI systems capable of being seriously misused. In this chapter, we illustrate this through case studies of scam compounds and their associated underground tool marketplaces, and online communities for the generation of synthetic child sexual abuse material. This poses a problem, as regulating devices is more difficult and in many ways more invasive than governing cloud services, given the private nature of activities on these devices and the more limited control and visibility rule-setters have over them. Despite this, building in limits on the software that can run on individuals' computers has a long history, stretching from the Nintendo Entertainment System in the mid-1980s to the iPhone today. We identify and characterize two approaches to regulating activities on devices: software gatekeeping, most familiar from the App Store, and operating system surveillance, which has been at the heart of debates on 'client-side scanning.' We consider how these might apply to the abuse of AI systems, examining the theoretical possibilities and limitations of approaches including notarization of models, restricting access to on-device machine learning hardware and behavioral monitoring. Given that extending these approaches to every device on Earth is unrealistic, we draw on other areas of technology policy where there have been successful efforts at segmenting 'safer' publicly available devices from more specialized open-ended ones. Careful combinations of market design and technical interventions could be deployed to offer a limited — if thorny and potentially dangerous — path forward to disrupting the most harmful forms of locally facilitated AI misuse.

FACE OFF

In January 2024, a Hong Kong-based clerk received an urgent email. They were working for Arup, a huge corporation best known for engineering some of the world's most esoteric structures, like the Sydney Opera House and the London Zoo's infamous Penguin Pool.[1] Rob Boardman, the Chief Financial Officer of the firm, had a pressing task for the clerk. He asked the staffer to join a conference call, in which it was made clear that there was a need to transfer a significant amount of money at short notice from accounts the clerk had access to. The meeting, also attended by various other senior Arup employees, was then followed by detailed instructions sent directly by the CFO. The clerk carried out these orders, transferring the equivalent of twenty-five million dollars in fifteen transactions to five bank accounts.[2]

1: The London Zoo Penguin Pool was closed to penguins in the early 2000s after it became clear that the spiraling, gravity-defying, modernist concrete walkways for the birds were scratching their feet and causing a bacterial infection called bumblefoot.

2: Cheng Leng and Chan Ho-him, "Arup Lost $25mn in Hong Kong Deepfake Video Conference Scam" (*Financial Times*, 16 May 2024) https://www.ft.com/content/b977e8d4-664c-4ae4-8a8e-eb93bdf785ea accessed 12 April 2025; Tang Ping-keung, "LCQ9: Combating Frauds Involving Deepfake (Response to a Question by Dr the Hon Tan Yueheng)" (*The Government of the Hong Kong Special Administrative Region*, 26 June 2024) https://www.info.gov.hk/gia/general/202406/26/P2024062600192.htm accessed 12 April 2025; Dan Milmo, "Company Worker in Hong Kong Pays out £20m in Deepfake Video Call Scam" (*The Guardian*, 5 February 2024) https://www.theguardian.com/world/2024/feb/05/hong-kong-company-

Upon filing a report to head office that these transfers had been completed, the clerk came quickly to realize what the reader may already have: the instructions were fake. The transfer was a scam. Neither Rob Boardman nor any of the senior colleagues he had brought onto the conference call were real: their likenesses had been generated and altered using AI tools based on publicly available videos. Given the hierarchy and stiffness of formal online meetings, the video seems likely to have been created in advance and played to the unwitting staffer via videoconferencing software, the scenario playing out until trust was established.[3]

The precise tools these scammers used to swap or puppet these faces are not known. But we do know where similar scammers find similar AI systems, and one governance issue is striking. These tools do not always use contemporary AI models in the cloud. Instead, they often run on scammers' own devices.

A SCAMMER'S TOOLKIT

Telegram is an app created by brothers Nikolai and Pavel Durov in 2013. The pair of Russian Internet entrepreneurs first created VKontakte (or VK, the 'Russian Facebook') in the mid-2000s before eventually falling out of favor with Vladimir Putin and losing control of the platform to a new ownership group of regime-friendly oligarchs.[4] Eager to bounce back quickly with a new venture, the Durovs established Telegram, launching it first in Russia and taking it global via legal entities in Dubai and the British Virgin Islands.

Telegram differs from many other messenger apps in its facilitation of flexible groups that can have hundreds of thousands of members. Underpinning this functionality are user-led infrastructural *bots*, which allow group administrators to automate posts, tailor content ranking, integrate payment systems and more, creating what in effect amounts to a platform-in-a-platform. Since the launch of their *Bot API* in 2015, the openness of Telegram to third-party code has spawned a sprawling network of functionality, which has become especially popular in certain communities and countries (in particular, Iran, where Telegram appears to have the majority of its global users, and where hundreds of thousands of massive groups are used daily for news, entertainment, political chat and more).[5]

As is typically the case with computational generativity, not all of the innovative uses of these bots have been unproblematic. Alongside Farsi news channels or German climate activist networks, Telegram bots are also popularly used to facilitate black markets for hard-to-get goods and services typically used by Internet-oriented criminal enterprises. The largest of these are the Mandarin-language megamarket Huione/Haowang Guarantee (汇旺/好旺担保, which shut down during our period of research) and the marketplace widely pointed to as its successor, Tudou Guarantee — literally, 'Potato Guarantee' (土豆担保). In these Telegram channels, sellers deposit cryptocurrency, typically the Tether 'stablecoin,' which, along with a system of escrow for the purchase of goods or services, acts as a 'guarantee' of vendor

deepfake-video-conference-call-scam accessed 12 April 2025.

3: Ping-keung (n 2).

4: Sergey Kozlovsky, "Here's Why Mail.Ru's Complete Control Over VKontakte Is Bad News" (*Global Voices*, 25 September 2014) https://globalvoices.org/2014/09/25/russia-vkontakte-mailru-kremlin-social-network/ accessed 6 January 2026.

5: Ahmed Al-Rawi, "News Loopholing: Telegram News as Portable Alternative Media" (2022) 5(1) Journal of Computational Social Science 949 DOI: 10.1007/s42001-021-00155-3 accessed 6 January 2026.

阿昌哥 (出售过仁脸技术)	Achang (Sold Human Face Technology)
蚂蚁 Ai 换脸-精聊直播专业版	Ant AI Face Swap - Pro Chat Live Version
熊猫 AI 换脸变声模真机竖屏	Panda AI Face Swap Voice Change Real Machine Vertical Screen
匿名者 AI 换脸声建模防识别	Anonymous AI Face Swap Sound Modeling Anti-Recognition
AI 换声-AI 换脸-建模	AI Voice Swap - AI Face Swap - Modeling
胖哥团队过一切 AI 仁脸验证	Pang Ge Team Pass All AI Human Face Verification
小米 AI 换脸+换声+建模	Xiaomi AI Face Swap + Voice Swap + Modeling
昊天 AI 换脸-独家核心技术	Haotian AI Face Swap - Exclusive Core Technology
AI 换脸大师 (精聊神器)	AI Face Swap Master (Pro Chat Tool)
单图换脸-实时换脸-无需建模	Single Image Face Swap - Real-time Face Swap - No Modeling Needed
辰信 Ai 科技-实时换脸-换声	Chenxin AI Technology - Real-time Face Swap - Voice Swap
AI 非一图产品防手机 AI 识别	AI Non-Single Image Product Anti-Mobile AI Recognition
(外星人 Ai 实时换脸)	(Alien AI Real-time Face Swap)
全行业 AI 仁脸 KYC 仁脸验证	All Industry AI Human Face KYC Human Face Verification
银行 ai 代过仁脸	Bank AI Proxy for Human Face
(英雄) 银行人脸技术操作	(Hero) Bank Face Technology Operation
小杰 Ai 过仁脸	Xiao Jie AI Pass Human Face
极速 Ai 团队代操作	Rapid AI Team Proxy Operation

Figure 3.1: AI face-swap groups on Huione Guarantee as of April 2025 (with English translation, right). Each group typically represents a discussion forum in which to inquire about or purchase a distinct product. Deposits refer to cryptocurrencies deposited with Huione Guarantee for credibility. Prepared by authors.

credibility and reliability.

Such guarantees and escrow services are important in a world where deception is not just common, but is itself the product. The opening pages of the Huione and Tudou platforms offer a list of seller categories: for instance, one can obtain the services of 'brushers' (招刷手) who inflate reviews for e-commerce products, buy artificial followers and traffic (收粉收量) or borrow Apple IDs (租苹果 id). These services for cybercriminals are part of the general industrialization and platformization of hacking, an evolution from scrappy go-it-alone script-kiddies to a global market for illicit infrastructure-as-a-service.[6]

For the story of the fake corporate executive, most relevant is the cutting edge of illicit AI face-swap software (AI 换脸). When navigating to this part of the marketplace, a user will see a large array of channels, each of which features discussion of a particular suite of AI face-swap software with tens of thousands of dollars' worth of cryptocurrency lodged per channel with the marketplace as a deposit and a marker of seller honesty (Figure 3.1). In a sea of potential offerings, the branding and marketing are evocative, with models given names like 'Hàotiān AI' (昊天), meaning 'vast heaven,' or 'huā fēi huā' (花非花) — "a flower, yet not a flower" — a reference to the illusory, dreamlike image conjured in a famous poem by Tang dynasty poet Bái Jūyì.[7]

On these illicit AI model markets, two core use cases quickly become apparent. The first involves the sale of software intended primarily for swapping faces and voices to fool or scam other people in live conversations. Sellers often showcase the capabilities of their models by impersonating famous celebrities. Voice manipulation models are also typically bundled into

6: Ben Collier and others, "Cybercrime Is (Often) Boring: Infrastructure and Alienation in a Deviant Subculture" (2021) 61(5) The British Journal of Criminology 1407 DOI: 10.1093/bjc/azab026.

7: Translated version in Xu Yuan Zhong (tr), *Songs of the Immortals: An Anthology of Classical Chinese Poetry* (Penguin 1994) 84. Chinese original at https://zh.wikisource.org/wiki/%E8%8A%B1%E9%9D%9E%E8%8A%B1.

the suite, often with the ability to submit media so that purchased models can be customized to a desired person's voice or face without buyers having to do it themselves. The second set of tools is customized in particular to fool 'know-your-customer' (KYC) identity verification requirements. Most commonly used in online banking, these KYC tools typically require one to show an identity document and then record a video with one's face.

Both use cases require high-quality generative AI systems. Sellers highlight the smoothness of their tools, as a glitchy face is a surefire way to create suspicion in a romance or business scam situation. KYC applications typically ask people to move their head, to open their mouth and sometimes to cover part of their face, at times involving a live human agent who compares the video to the submitted identity documents. Most videos of these systems in action are at pains to demonstrate how their models continue to function in all those situations and more (Figure 3.2). One provider states:

> *During a video you can blow a kiss for the client; touch facial features; pinch the nose; pinch the face; head turns with clear edges; perfectly resist known AI recognition tech; more realistic... a lot of low-quality AI tech is instantly spotted by clients and can't handle touching the face, wearing glasses, walking, shaking, full-screen software, etc.*[8]

These technologies are powered by a mix of open-source and proprietary components. Several openly state their reliance on open-source frameworks as a base, such as the Retrieval-based Voice Conversion system (RVC), a community-built framework for speech-to-speech conversion.[9] Some even decry 'unscrupulous resellers' (奸商) of open-source tools, emphasizing that they, the reputable sellers, help users download free software and provide significant added value on top of it: user interfaces, systems to inject video input into mainstream apps, detection-avoidance technologies and more. This is important as, to work properly, scam-optimized AI systems need to overcome various safeguards. Sellers boast of being able to avoid the countermeasures increasingly deployed by messaging apps such as WhatsApp, Facebook Messenger and LINE to block video feeds from being piped in from sources other than the on-device camera. Some also claim to be able to dodge AI detection tools built at the hardware level into some modern smartphones, particularly from Chinese manufacturers like Huawei. These protective systems are discussed further in Chapter 6.

Running this kind of criminal-grade software generally requires a setup significantly more powerful than the average consumer device. The recommended setup described by one Huione seller is a desktop with an Intel 13th Generation Core i9-13900 — one of the highest-tier CPU chips Intel sells — and an Nvidia GeForce RTX 4090, an enthusiast-level GPU that launched in late 2022. Those two components alone cost between $2,000 and $3,000, and the rest of the build would require much more memory than an ordinary desktop as well as a water-cooling system: the type of computer typically owned by a video editor, 3D graphic designer or extremely serious gamer.[10] A major benefit of running all these technologies lo-

8: Quote from authors' research into Huione Guarantee; machine translation from Mandarin. Original data on file with authors.

9: RVC-Project, "RVC-Project/Retrieval-based-Voice-Conversion-WebUI" (*GitHub*, 2025) https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI accessed 12 August 2025.

10: These kinds of models may foreseeably run on cheaper, ordinary devices in the near future, as video models become more efficient and leading consumer devices integrate specialized hardware capable of applying machine learning to live data streams, such as for "live translation" features.



Figure 3.2: Sales video from Huione Guarantee for the 'Haotian AI' face-swap product, demonstrating (as Elon Musk) the ability of users to place their hands on their face, and demonstrating a gender face-swap with accompanying wig. (These models do not currently appear to be able to swap hair as well.) The right-most individual is the user without face modifications. Prepared by authors.



Figure 3.3: 'KK Park,' a human-trafficking and Internet fraud hub located in Myanmar, at the border with Thailand (16°38′ 51.2"N 98°31′ 14.6"E). Source: Google Maps (retrieved April 2025).

cally is the lack of detectability. Scammers are aware that their usage is being tracked, with interview studies noting that those involved in these scams are regularly given strict instructions to avoid Chinese cloud AI services due to the risk of surveillance and disruption.[11]

SCAM COMPOUNDS

Many users of these tools are likely to be found in *scam compounds* that are emerging around the world. These are physical hubs, centers of operation for international organized Internet criminals, typically located in places that are difficult for law enforcement to access. Southeast Asia has emerged as a hotspot for these compounds. Amnesty International in 2025 identified

11: Gilad Gressel and others, "Love, Lies, and Language Models: Investigating AI's Role in Romance-Baiting Scams" in (USENIX Security '26, 2026) DOI: 10.48550/arXiv.2512.16280.

53 in Cambodia alone.[12] Cambodia, incidentally, is the country most closely associated with the running of both Huione and Tudou Guarantee.

KK Park is one of the most notorious scam compounds. Tucked into a meander of the Moei River separating Myanmar and Thailand, it is a sprawling, franchise-style criminal hub, with floors and buildings of variously connected illegal Internet operations that share and swap resources and humans (Figure 3.3). Many of them are not there of their own volition. Scam compounds are notorious destinations for human traffickers, with vulnerable, skilled individuals making their way across the world with the promise of a legitimate technology-related job, only to find themselves imprisoned and forced to carry out illegal activities under duress.[13] With their passports taken away, surrounded by guards and often complicit local officials, escape is extremely difficult.[14]

These individuals are organized in rooms and floors according to the type of scams being undertaken, with labor divided according to skill and to the equipment they need to carry out these scams. Consider romance scams, which are now often called 'pig butchering' scams after the Chinese term for them, shāzhūpán (杀猪盘). The process is a multi-stage one, which involves 'finding pigs' (zhǎo zhū, 找猪), such as from social networks or marketing agencies; 'feeding pigs' (wèi zhū, 喂猪), by romancing them, sending them gifts and similar; a 'pig-raising period' (yǎngzhū shíqī, 养猪时期), where they are introduced to the heart of the scam, for instance by persuading them to join a realistic investment platform where they see their money or cryptocurrency grow (and they may even be able to withdraw some of their provided funds to boost their confidence in the system's authenticity); and finally 'slaughter time' (shāzhū shíkè, 杀猪时刻), where they are locked out of the system, ghosted and see their funds disappear forever. A team of different people may operate different stages of this scam on a single individual.[15]

Those trafficked to scam compounds often have to bargain for their freedom, typically by recruiting replacements, or by getting family or friends to pay an enormous ransom. Not everyone is given that option, though. In 2025 *The Economist* reported that "anyone with any kind of knowledge of AI" now finds it much more difficult to get out.[16] They are simply too valuable.

The growing value of AI practitioners to scammers is evident in the rising number of media reports of scammers using AI: for instance, generating realistic-sounding imitations of people's family members asking for money.[17] Chinese documentary filmmaker and journalist "Hua

12: Amnesty International, ""I Was Someone Else's Property": Slavery, Human Trafficking and Torture in Cambodia's Scamming Compounds" (*Amnesty International*, 2025) https://www.amnesty.org/en/wp-content/uploads/2025/06/ASA2394472025ENGLISH.pdf accessed 13 August 2025.

13: Although contrasting accounts have argued that some participants are staying willingly. See for instance Wu Dong, Wang Sisi, and Du Jingzhao, "一个记录者眼中，真实的缅甸电诈、危险与人性 (In the Eyes of a Chronicler: The True Myanmar Telecom Fraud, Danger, and Human Nature)" (冷杉 *RECORD (*凤凰 *WEEKLY)*, 17 January 2025) https://perma.cc/BWN8-RN8P accessed 13 January 2026.

14: Ivan Franceschini, Ling Li, and Mark Bo, *Scam: Inside Southeast Asia's Cybercrime Compounds* (Verso Books 2025).

15: ibid.

16: Sue-Lin Wong, "Scam Inc.: 3. The Bottom Line" (*The Economist (podcast)*, 6 February 2025) https://www.economist.com/podcasts/2025/02/06/3-the-bottom-line accessed 20 February 2025, 16.02 minutes.

17: Shao Luwen and Wen Jinghua, "你的声音被谁 "偷" 走了？——AI 声音滥用现象调查 (Who "Stole" Your Voice? — Investigation into the Abuse of AI Voices)" (*ST Daily*, 12 February 2025) https://www.stdaily.com/web/gdxw/2025-02/12/content_296148.html accessed 18 October 2025; Charles Bethea, "The Terrifying AI Scam That Uses Your Loved One's Voice" (*The New Yorker*, 7 March 2024) https://www.newyorker.com/science/annals-of-artificial-intelligence/the-terrifying-ai-scam-

Zong" (real name Wu Dong), who has extensively covered and visited scam compounds, has claimed that the search for AI talent has led to scammers operating "Masters squads" and "PhD squads" using advanced techniques to entrap the highest-value marks.[18] Such a squad may have been behind the sophisticated Arup fraud described earlier.

Less widely covered, however, are the ways in which AI systems are being embedded into the day-to-day process of 'pig butchering.' All scammers in a compound are explicitly taught to use language models to clean up their language and improve the sentiment and alluring nature of their messaging.[19] A 2026 study interviewing over a hundred scam insiders from 2022 to 2025 found that in its final two years, all 34 interviewees mentioned the use of AI in their daily operations.[20] Generative models are reportedly widely used by scammers to generate conversations with the aim of largely or even fully automating manipulative romantic interactions with limited human oversight.[21] Some malicious AI providers are not shy about the fact that their tools will be used in scam compounds. Hàotiān AI's website offers "on-site installation in parts of Cambodia."[22] Romance scammers have explained some of their current practices to researchers, and these go beyond just live-video and audio manipulation, also including text generation and interactions with targets: for instance, wiring up language models to Apple's Siri system to automatically respond to incoming messages with a 'perfect response.'[23] Scammers also seek out and hire 'AI face models,' who can, by combining their appearance, tone and accent with AI alterations, tip the scales in a victim's mind from unconvincing to realistic enough that they part with their cash.[24]

A world of autonomous, 'agentic' systems let loose to try to scam individuals in the best way they see fit remains speculative for now, and yet increasingly seems not so far away. Even if generative scam systems are brittle and imperfect, scammers can afford for these systems to be very error-prone: unlike a *bona fide* firm deploying such a system to interact with their customers, scammers care little about the average user's experience — at scale, even a small success rate can yield lucrative rewards.

that-uses-your-loved-ones-voice accessed 18 October 2025.

18: Dong, Sisi, and Jingzhao (n 13).

19: Andy Greenberg, Lily Hay Newman, and Matt Burgess, "He Leaked the Secrets of a Southeast Asian Scam Compound. Then He Had to Get Out Alive" (*WIRED*, 27 January 2026) https://www.wired.com/story/he-leaked-the-secrets-southeast-asian-scam-compound-then-had-to-get-out-alive/ accessed 31 January 2026.

20: Gressel and others (n 11).

21: Simon Moseley, "Automating Deception: AI's Evolving Role in Romance Fraud" (*Centre for Emerging Technology and Security (CeTaS)*, 15 April 2025) https://cetas.turing.ac.uk/publications/automating-deception-ais-evolving-role-romance-fraud accessed 15 April 2025; Nick McKenzie, "Undercover Investigations: How AI Is Supercharging Romance Scams" (*Netskope*, 30 July 2025) https://www.netskope.com/blog/undercover-investigations-how-ai-is-supercharging-romance-scams accessed 18 October 2025.

22: See archive of the Hàotiān AI website at https://web.archive.org/web/20240524155426/https://www.haotian.ai/ ("以及我们还提供柬埔寨部分地区的上门安装服务或是远程便捷安装"); see generally on this firm Matt Burgess, Lily Hay Newman, and Zeyi Yang, "The Ultra-Realistic AI Face Swapping Platform Driving Romance Scams" (*Wired*, 20 December 2025) https://www.wired.com/story/the-ultra-realistic-ai-face-swapping-platform-driving-romance-scams/ accessed 20 December 2025.

23: Jonathan Barnor and Sheena Lovia Boateng, "The Romance Ruse: Exploring Dark Entrepreneurship in AI-Driven Fraudulent Activities" in Sheena Lovia Boateng and Richard Boateng, *AI and Society: Navigating Policy, Ethics, and Innovation in a Transforming World* (Productivity Press 2025) DOI: 10.4324/9781003595670 59.

24: Matt Burgess, "'100 Video Calls Per Day': Models Are Applying to Be the Face of AI Scams" (*Wired*, 16 March 2026) https://www.wired.com/story/models-are-applying-to-be-the-face-of-ai-scams/ accessed 6 August 2026.

SYNTHESIZED CHILD SEXUAL ABUSE

The second troubling form of local generative AI misuse involves child sexual abuse material, once commonly termed 'child pornography.' The more technically savvy child sexual abusers congregate on forums on the darkweb — the popular name for "hidden services" only accessible through the Tor network (and designed to mask their server location).[25] Hidden services are characterized by their whimsical .ONION domain ending, and are far from universally illicit — a significant majority of the known domains are simply alternative ways to access publicly available services, including ways for individuals in oppressive regimes to access major technology firms' services, or for whistleblowers to drop material anonymously to journalists.[26] The U.S. government has, since the network's inception, been a major funder of Tor, with the State Department providing several million dollars a year toward the project.[27] These forums are studied by academics, NGOs and law enforcement, typically using tools which scrape only the sites' text to avoid encountering illegal content. Such studies provide relevant insights into the modalities of contemporary AI use in these communities.

Abusers use open-source, on-device models as their "tool of choice" because of the limited moderation and detection capabilities possible when cloud intermediaries are not involved.[28] These are often simply major open-source image models (e.g. Stable Diffusion, Flux) wrapped in widely available user-friendly interfaces.

In 2023, the "overwhelming consensus" on darkweb forums was that AUTOMATIC1111's WebUI tool was the simplest and best documented of these interfaces.[29] This tool is not hidden, far from unique and not only used for abusive purposes. When open-source model providers release a model, communities, including artists, technology companies and hobbyists, build applications such as AUTOMATIC1111 that allow these models to be easily used. Similar options include open-source hobby tools (e.g. ComfyUI, Forge), tools aimed at researchers (e.g. Hugging Face Diffusers), tools produced by industry (e.g. Apple's Core ML Stable Diffusion) or tools aimed at consumers on marketplaces like Apple's App Store (e.g. Draw Things).

Mainstream models often display innately harmful abilities that are hard to control,[30] may

25: Ian Levy and Crispin Robinson, "Thoughts on Child Safety on Commodity Platforms" (*arXiv*, 19 July 2022) DOI: 10.48550/arXiv.2207.09506 21.

26: Alfonso Rodriguez Barredo-Valenzuela, Sergio Pastrana, and Guillermo Suarez-Tangil, "Snorkeling in Dark Waters: A Longitudinal Surface Exploration of Unique Tor Hidden Services" (2025) 20 IEEE Transactions on Information Forensics and Security 5386 DOI: 10.1109/TIFS.2025.3573662.

27: CohnReznik LLP, "Financial Statements and Independent Auditor's Report" (*The Tor Project, Inc*, 30 June 2024) https://www.torproject.org/static/findoc/2023-2024-The_Tor_Project-AuditedFinancialStatements.pdf accessed 19 October 2025.

28: Internet Watch Foundation, "What Has Changed in the AI CSAM Landscape?" (*IWF*, July 2024) https://www.iwf.org.uk/media/nadlcb1z/iwf-ai-csam-report_update-public-jul24v13.pdf accessed 15 April 2025, 14; Internet Watch Foundation, "How AI Is Being Abused to Create Child Sexual Abuse Imagery" (*IWF*, October 2023) https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf accessed 15 April 2025, 45.

29: Internet Watch Foundation, "How AI Is Being Abused to Create Child Sexual Abuse Imagery" (n 28) 22. Note that while the IWF tried to redact the name of the tool here, they only redacted the GitHub user's name (AUTOMATIC1111), not the tool name, WebUI, making it clear which tool is being talked about. In any case, a similar finding is made in Deanna Davy and Samantha Lundrigan, "Artificial Intelligence-Produced Child Sexual Abuse Material: Insights from Dark Web Forum Posts" (*International Policing and Public Protection Research Institute, Anglia Ruskin University*, July 2024) https://www.aru.ac.uk/-/media/Files/ipppri/IPPPRI-Insight-No-1-AI-CSAM.pdf accessed 13 August 2024.

30: See generally Chapter 2.

be inadvertently trained on child sexual abuse material scraped from the Web,[31] and appear difficult to censor due to the open-ended nature of text and the ability to chain innocuous prompts into harmful ones.[32] This means that standard, off-the-shelf image models, without even requiring specific CSAM-oriented fine-tuning, can generate realistic sexualized content depicting minors, if not outright CSAM. This was made publicly evident in December 2025, when xAI's Grok model, which can be prompted by any user on Elon Musk's X platform, was queried to produce large quantities of 'undressed' images of celebrities and children.[33] Most providers are more diligent in creating input and output filters to prevent this kind of activity from happening (see Chapter 2). But through interfaces like AUTOMATIC1111, local models customized to produce deepfakes of certain individuals (or to create more realistic pornography) can be downloaded from various model hosting services (we will discuss these further in Chapter 5) and run on a powerful laptop or desktop device without detection.

ABLITERATION

To generate the kinds of content they want, online communities are finding ways to remove protections that have been baked into many downloadable genAI models. In 2024, a collaboration of researchers from ETH Zurich, Anthropic, MIT and Maryland yielded an interesting discovery: there appeared to be a simple way to, after a model was released, remove a key safeguard built in by developers during training — that the model, when queried with some kinds of impermissible requests, refuses to generate the requested content.[34] The insight from this paper came from the authors' conceptualization of how one might think about the process of concept formation or association inside generative models, and the ways in which this can create weak points within them to be attacked. We will simplify their argument for illustrative purposes.

Imagine that you ask an LLM to provide you with detailed instructions on how to make a bomb. When receiving the token "bomb," the layers of a model that trigger on similarly used words — grenade, dynamite and other *types of weapons* — might engage. Given the association that bomb has with other *types of weapons* (rocket, missile, cannon) in large text corpora, intermediate layers that trigger when *very dangerous things* (not just bombs, but perhaps also knives, guns or even more abstract things that can kill, such as viruses) are discussed might then engage. This might in turn be associated with the higher parts of the model that have been trained to *refuse to answer* on such topics.

However, what if the *refuse to answer* constellation inside the model is the same or similar for all topics that a model refuses to talk about? Might it be possible to isolate and delete this

31: David Thiel, Melissa Stroebel, and Rebecca Portnoff, "Generative ML and CSAM: Implications and Mitigations" (*Stanford Internet Observatory, Cyber Policy Center*, 24 June 2023) https://doi.org/10.25740/jv206yg3793.

32: David Glukhov and others, "Position: Fundamental Limitations of LLM Censorship Necessitate New Approaches" in *Proceedings of the 41st International Conference on Machine Learning* (International Conference on Machine Learning, PMLR 2024) https://proceedings.mlr.press/v235/glukhov24a.html.

33: Justin Hendrix and Riana Pfefferkorn, "The Policy Implications of Grok's 'Mass Digital Undressing Spree'" (*Tech Policy Press*, 4 January 2026) https://www.techpolicy.press/the-policy-implications-of-groks-mass-digital-undressing-spree/ accessed 6 January 2026.

34: Andy Arditi and others, "Refusal in Language Models Is Mediated by a Single Direction" in *Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)* (The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024) DOI: 10.48550/arXiv.2406.11717.

part of the model, such that there is no concept of *refuse to answer* to activate? That was the core proposition of the paper.

After a pre-print of 'Refusal in Language Models Is Mediated by a Single Direction' was posted to the Web, online communities jumped onto it. They quickly coined a neologism — **abliteration**, from a combination of *ablation*, the removal of bad tissue, and *obliteration* — and set to work trying to entirely stop AI models from ever refusing their queries.[35] Users of these customized models would be able to have all of the functionality of the original, without worries about model guardrails that required prompt engineering or other techniques. You could simply ask for what you wanted. 'Abliterated' versions of major language models, including LoRAs which achieved the same or similar effects, quickly appeared on major hosting platforms such as Hugging Face, accompanied by tutorials about how to use or create them. Over 11,000 abliterated or similarly altered "uncensored" language models are available on Hugging Face as of 2025, with a total of 19 million downloads across them.[36]

Researchers have been developing defenses against this technique, but they can come with costs, such as a reduction in the general performance of the model.[37] In general, such "white box" attacks, where the attacker has full access to the model, are difficult to secure against in robust and meaningful ways, leading to cat-and-mouse situations between developers and users. In any case, even where such defenses exist, meaningful safety would require them to be adopted by all publishers of open models, which presents a serious coordination problem, especially where not adopting some defenses gives model providers other competitive advantages in performance.

ADAPTORS FOR ABUSE

However, while many mainstream models, particularly if 'abliterated' to remove their propensity to refuse, are capable of generating imagery that might legally be categorized as child sexual abuse material, that does not mean that this type of imagery is of a quality to be gratifying to those seeking this material.[38] As a poster on a darkweb child sexual abuse image sharing forum studied by academics put it:

> *Getting back to AI images: I was fooled initially by one and then once I learned it was AI generated I zoomed in, looked carefully, etc and then of course I noticed the fake features that were otherwise very subtle (skin too smooth, teeth too perfect etc). And my attraction or hornyness if you like dropped like a hammer.*[39]

Guardrails placed on mainstream models at the model layer might not prevent imagery qualifying as CSAM from being created when someone is determined to circumvent them,

35: FailSpai, "Abliterated-v3: Details about the Methodology, FAQ, Source Code; New Phi-3-mini-128k and Phi-3-vision-128k, Re-Abliterated Llama-3-70B-Instruct, and New "Geminified" Model." (*r/LocalLLaMA*, 28 May 2024) https://www.reddit.com/r/LocalLLaMA/comments/1d2vdnf/abliteratedv3_details_about_the_methodology_faq/ accessed 25 August 2025.

36: Zilong Lin and others, "Consiglieres in the Shadow: Understanding the Use of Uncensored Large Language Models in Cybercrimes" (*arXiv*, 18 August 2025) DOI: 10.48550/arXiv.2508.12622.

37: Harethah Abu Shairah and others, "An Embarrassingly Simple Defense Against LLM Abliteration Attacks" (*arXiv*, 25 May 2025) DOI: 10.48550/arXiv.2505.19056.

38: Internet Watch Foundation, "What Has Changed in the AI CSAM Landscape?" (n 28).

39: Davy and Lundrigan (n 29) 4.

but such precautions can limit the quality and utility of any illicit content to abusers. The most obvious 'trust and safety' reaction to potential CSAM generation would involve the removal of risky imagery of children from the training data, or, going even further, the removal of legal pornographic material more generally. Models may still learn these concepts given that sexuality is omnipresent in human life (and therefore, also in AI training data, which can feature movie screengrabs, art, etc.), but data cleaning may reduce the capability to produce CSAM desirable to abusers.

When these kinds of governance techniques fully or partially succeed, it means that dual-use capabilities of a model will be limited unless they are augmented with further *adaptors*. As another illicit forum participant put it:

> *It seems to me that this kind of sensitive data like realistic anatomy of child/adolescent body and genitalia will never appear in publicly available checkpoints. So definitely a good LoRA will always be needed for our purposes.*[40]

What is a 'LoRA'? Generally, genAI fine-tuning involves taking a model, exposing it to more data using a training algorithm and, in doing so, updating the weights in what is called a new "checkpoint." For instance, by training an open-weight model on a corpus of CSAM, one can alter it to depict material that was deliberately omitted during its training, or that it was later fine-tuned not to generate. This process requires significant compute capabilities. Classic forms of fine-tuning are much more demanding than simply querying a model and receiving outputs, in effect amounting to a process of retraining which is difficult or impossible to do on typical consumer hardware. Part of the problem is that the resulting model is also typically just as large as the one you started with — tens or hundreds of gigabytes — and one needs high-performance GPUs. Abusers wishing to fine-tune models this way may have been forced to do so via cloud environments, and thus be subject to the potential governance mechanisms outlined in Chapter 2.

However, in 2021, researchers from Microsoft created a method to fine-tune in a way that does not require nearly as much computing or storage capacity.[41] At its core was a new technique to map between the millions, billions or even trillions of weights in a contemporary AI model (computationally expensive to change) and a smaller set of parameters that are linked to the model weights in a one-to-many relationship. To understand how this works, imagine a matrix with as many datapoints as there are parameters in the model itself (Figure 3.4, left). There are many ways to multiply two much, much smaller matrices together to make a matrix with the same dimensions as the big one (Figure 3.4, right). To put it in even simpler and more analogous terms (particularly if you are wondering what matrix multiplication is), ten times ten is one hundred, but ten plus ten is twenty. These much smaller matrices can be the things that are instead somewhat 'fine-tuned' via a much less intensive course of training. The end result is a statistical representation that doesn't contain the original model itself, but

40: Emmanouela Kokolaki and Paraskevi Fragopoulou, "Unveiling AI's Threats to Child Protection: Regulatory Efforts to Criminalize AI-Generated CSAM and Emerging Children's Rights Violations" (*arXiv*, 1 March 2025) DOI: 10.48550/arXiv.2503.00433 11.

41: Edward J Hu and others, "LoRA: Low-Rank Adaptation of Large Language Models" in *International Conference on Learning Representations (ICLR)* (International Conference on Learning Representations, 2021) DOI: 10.48550/arXiv.2106.09685.

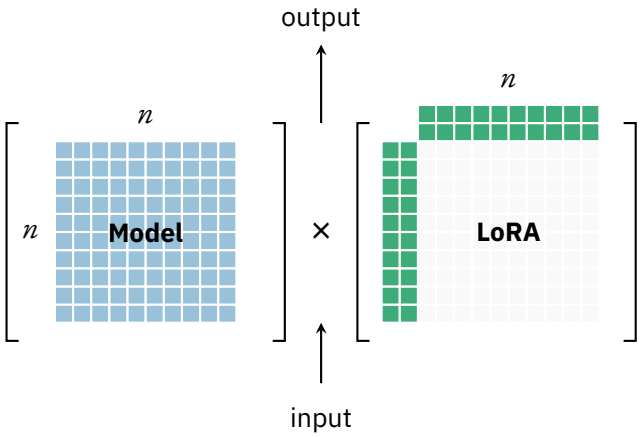

Figure 3.4: Schematic of how a LoRA works. The untouched weights of the model on the left are multiplied by the LoRA. The LoRA itself is not the same size as the model, but is made of two separate components which when multiplied together make a matrix the same size as the parameters in the main model. As a result, many fewer parameters need to be trained — indicatively, only the shaded squares on the right. Simplified for communication. Diagram by authors.

rather a smaller matrix that, once expanded, is exactly the same size as the original model (the big light square bordered by the darker weights in Figure 3.4, right).

This technique, *low-rank adaptation* or LoRA for short, also sometimes called 'parameter-efficient fine-tuning,' is not as precise as traditional fine-tuning, because of the one-to-many mapping of the LoRA parameters. Nonetheless, it can prove highly effective in practice. These, along with other related methods to produce plug-ins that alter models after training, are now commonly referred to as **adaptors**. These adaptors deploy math that can be wielded to shape a generative model's parameters like a filter or lens, altering them in new and subtle ways.

The large online communities tinkering with genAI models quickly adopted LoRAs. In malicious user circles, LoRAs have been trained on a mix of abusive and non-abusive data to achieve their ends, such as real CSAM and abuse images, as well as publicly available images of famous children:[42]

> *I'll start by saying I though I was a pretty technical person. But this AI stuff has overtaken me. I'm trying to even get familiar with the terminology - Loras, Stability, Stable Diffusion ... My goal: I have a bunch of images of a girl from her Instagram, maybe 20-30. I want to create CP [child pornography] of her, using her face as a training model.*[43]

> *Without LoRA, the differences in anatomy between the checkpoints are more noticeable.*[44]

42: Kokolaki and Fragopoulou (n 40).
43: Davy and Lundrigan (n 29) 4.
44: Kokolaki and Fragopoulou (n 40) 10.

Fine-tuned LoRAs trained to produce the likenesses of celebrities are also widely available, in particular relatively minor, and usually female, Internet celebrities.[45] Paired with permissive model checkpoints or chained with additional LoRAs, models trivially produce image-based sexual abuse of these individuals.[46] When users add dedicated child-abuse LoRAs, such models can produce "de-aged" images resembling these individuals, adding a new layer to image-based sexual abuse.[47]

LORA LEGISLATION

The law in some jurisdictions has already adapted to the rise of such malicious image models. The creation, possession and distribution of child sexual abuse image generators, access to services providing them, or instructions to make or use them, are now criminal offenses in the UK.[48] But enforcing such legislation may be difficult. For example, let us assume that law enforcement agencies locate models or LoRA files in circulation that they suspect to be illicit. The problem is that, for both fine-tuned models and LoRAs, in the absence of specific changelogs or public instructions as to their use, it is technically quite difficult to determine what these adaptors actually change about the underlying model. Given that many LoRAs are benign and legal — for instance, improving the ability of image models to render a particular place or product — it may be difficult to gather evidence to prosecute in some situations.

The matter is more difficult given that the true, malicious purpose of a LoRA can be obscured. LoRAs which appear to enhance certain functionality can have malicious capabilities covertly integrated within them.[49] LoRAs typically come with one or more keywords or phrases that a model is trained to respond to in order to trigger the effect of the additional parameters they contain. It is possible that one set of words might trigger an innocent filter, while another might trigger child sexual abuse functionality. Without documentation, and without the right type of model to connect the LoRA to, these files are illegible and meaningless for investigators.

Even where both model and LoRA are present, our ability to understand complex machine learning models based on examining their internals — these days called *mechanistic interpretability* — is limited, and such work is slow and difficult. The statistical associations made by models often do not come from the same, or even similar, representations of the world as humans; to return to an example from earlier in this chapter, a token like 'bomb' could become associated with all sorts of different tokens that aren't necessarily conceptually linked

45: 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; 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.

46: Emanuel Maiberg, "Inside the AI Porn Marketplace Where Everything and Everyone Is for Sale" (*404 Media*, August 2023) https://www.404media.co/inside-the-ai-porn-marketplace-where-everything-and-everyone-is-for-sale/ accessed 24 August 2023.

47: Internet Watch Foundation, "What Has Changed in the AI CSAM Landscape?" (n 28) 17.

48: Sexual Offences Act 2003, ss 45A, 69 (as inserted by the Crime and Policing Act 2026).

49: See, by analogy, Hongyi Liu and others, "LoRATK: LoRA Once, Backdoor Everywhere in the Share-and-Play Ecosystem" (*arXiv*, 30 April 2025) DOI: 10.48550/arXiv.2403.00108.

in our understanding. Pixel groupings associated with certain human anatomical representations in images might also be associated, in a model's architecture, with pixel groupings that to us seem to have little to nothing in common.[50] Even the most highly resourced industry labs struggle to understand the different components of cutting-edge models and how they meaningfully link to functionality.

In other words, while some LoRAs are going to be illegal, and various governance stakeholders across the stack will seek to prevent their distribution, it is difficult to tell from the LoRA alone whether it was in fact trained on illegal material to do illegal things. Litigators and regulators may be able to glean important information from the context, and from metadata associated with a fine-tuned model and how it is described by the actors in question. But once actors act to cover their tracks, this will pose a serious challenge, especially given the high standards of proof required in criminal proceedings.

PROSPECTIVE MODES OF LOCAL CONTROL

To sum up the first half of this chapter: local models that are far from the 'frontier' systems typically at the heart of today's AI governance and 'AI safety' conversations can be, and are, easily repurposed for a range of socially damaging and abusive ends. Worryingly, robust safeguards against these types of misuse are difficult or impossible to bake in — even if all model providers agreed to implement them universally.

This leaves us in a tricky predicament. If one cannot fundamentally change the functionality and architecture of generative AI systems to durably prevent the worst forms of harm, is the only viable option remaining for policymakers to impose limits on how these systems can be deployed in the real world?

The desire of public and private actors to control individuals' devices in order to prevent unsanctioned conduct is hardly a new one in the history of computing or of technology governance more broadly. Sometimes the goals motivating these kinds of policy have been aligned with the topic of this book — interventions taken with the stated aim of preventing forms of illicit and damaging misuse. Other times, the actor interests at play are even more ambiguous and self-serving: for instance, limits on computation intended to help firms maintain their brand image or restrict avenues for unwanted competition.

We group these efforts at control over computing devices into two broad technical areas. The first, *software gatekeeping*, involves the role of local devices and the governance mechanisms that determine how they function — from their chips to their operating systems to their integration with wider networked environments. Software gatekeeping is, at its core, about what kind of software can or cannot run on a device. Although this saga played out initially in the area of intellectual property and rights over media content, especially in the video game and movie industries, it later became enmeshed in a more multifaceted way with platform business models and the changing nature of the distribution and monetization of software.

50: See generally on mechanistic interpretability Lee Sharkey and others, "Open Problems in Mechanistic Interpretability" (*arXiv*, 27 January 2025) DOI: 10.48550/arXiv.2501.16496; Naomi Saphra and Sarah Wiegreffe, "Mechanistic?" in Yonatan Belinkov and others (eds), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP* (Association for Computational Linguistics 2024) DOI: 10.18653/v1/2024.blackboxnlp-1.30.

A second relevant area is *operating system surveillance*. This is a newer area of prospective control, driven by the ways in which software development has become more and more networked in recent years. Effectively, these are various practices where your operating system monitors how you use your computer, either voluntarily, as a productivity boon, or — as is currently being heatedly debated in Europe under pressure from law enforcement — to combat the spread of certain forms of illegal content.

These two areas have yet to be comprehensively applied in an effort to limit the use of malicious AI systems. But they're a likely next frontier for AI governance. To what extent are they appropriate for limiting the misuse of models? And what are their trade-offs and accompanying costs?

SOFTWARE GATEKEEPING

Software gatekeeping is a mode of technoregulation. In its purest form, technoregulation attempts to establish a rule or policy that cannot be broken, because it will be enforced automatically by the technology itself. This is not just an online phenomenon: think of speed bumps on the road, or speed limiters built into some cars. In today's connected world, these have become increasingly prevalent. For instance, rental e-bikes in several London boroughs have technoregulation stipulated by contract, requiring firms to use GPS to limit their speeds to a crawl in certain 'geofenced' areas, and prohibiting rides from being concluded (or charging for ending them) outside a valid parking bay.[51] Enforcement is automatic and *ex ante*. Legislators often turn to mandating technoregulatory mechanisms where costs of monitoring and enforcement are high, such as where there are few barriers to breaking the law, and the ability to do so is distributed widely.

Technoregulation in the digital economy is tangled up with the concept of the *walled garden*. This flowery term stems from the telecommunications industry in the 1990s. Providers of early online services such as AOL, Prodigy and CompuServe — known at the time as the "Big Three" information services — sought to avoid the public Internet and instead sell access to infrastructure and content simultaneously.[52] These efforts ultimately failed. Today, few Internet service providers attempt to tie their services to exclusive content, and network neutrality laws secure access to the open Internet for users.[53]

But far from being the end of the walled garden, the concept flourished in a different direction, and now sits behind many of the most powerful platform businesses. Telecoms firms had indeed failed to overcome the design features of the Internet that were conceived to resist singular capture. But modern platforms instead employed a more sophisticated mix of law and technoregulation to enclose networks that might otherwise resist their efforts.[54]

51: London boroughs do not currently have direct powers in public law to require this, and instead regulate through private law by using contractual agreements.

52: Christopher T Marsden, "Property Rights in the Broadband Space" (2001) 6 International Journal of Communications Law and Policy https://ciaotest.cc.columbia.edu/olj/ijclp/ijclp_6/ijclp_6n.pdf.

53: Tim Wu, *The Master Switch: The Rise and Fall of Information Empires* (Knopf Doubleday Publishing Group 2010) (on the failure of AOL); see generally on net neutrality, Christopher T Marsden, *Network Neutrality: From Policy to Law to Regulation* (Manchester University Press 2017).

54: Malte Ziewitz and Ian Brown, "A Prehistory of Internet Governance" in Ian Brown (ed), *Research Handbook on Governance of the Internet* (Edward Elgar Publishing 2013) DOI: 10.4337/9781849805049 (on the design history of the Internet); Julie

Instead of trying to counter the design features of the underlying network, these new actors rode upon them, constructing their own infrastructures on top — search engines, social networks, mobile operating systems and cloud services. They grew fast by initially placing low walls around their networks. Facebook lured users with a very large variety of third-party plug-ins; Google's search engine would index sites very permissively; Apple's iPhone would allow users to interact with 'Web apps' that did not require an app store to obtain. Twitter could be browsed easily using many unofficial clients, like TweetDeck, that had navigation, customization and automation features users found attractive.[55] However, once these platform effects had been exerted — relationships with various third parties built, a large userbase amassed — the walls could be raised. Platforms began capturing more and more value, locking in users through design decisions, monetization decisions and various technoregulatory interventions.[56]

POWER GAMES

Personal devices followed a similar trajectory. Early personal computers were typically very open and placed few restrictions on the kind of software that an individual could create for them or run on them. Device manufacturers soon began to build walls, often either to sell their own software or to create a platform where others might securely sell their products with less concern that they would be pirated.

Gaming consoles were among the earliest devices that were designed to run varied software in an extremely closed manner. They were often the first computers brought into people's homes by non-hobbyists. However, 1980s home consoles, notably the Atari 2600, were plagued by hastily thrown-together, poor-quality games. The 1982 release of the movie tie-in game *E.T. the Extra-Terrestrial* was a tipping point in this regard: unrealistic commercial expectations and consumer dissatisfaction led to 3.5 million of 5 million copies going unsold or being returned, leaving Atari on the hook for exorbitantly expensive licensing rights.[57] The video game is frequently derided as the worst of all time. The problem didn't stop there: only a year after E.T.'s release, 13 trucks were driven to a landfill in Alamogordo, New Mexico, to dump hundreds of thousands of game cartridges, including *E.T.*; countless thousands more were buried near the Atari HQ in California.[58] Poor-quality games and the resulting wob-

E Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism* (Oxford University Press 2019) DOI: 10.1093/oso/9780190246693.001.0001 (on enclosure and platform strategies); Carl Shapiro and Hal R Varian, *Information Rules: A Strategic Guide to the Network Economy* (Harvard Business School Press 1999) (a guide on how to lock users in and avoid the pitfalls of low network effects discovered by early Internet service providers like AOL; written by a later chief economist at Google).

55: Robert Gorwa and Douglas Guilbeault, "Unpacking the Social Media Bot: A Typology to Guide Research and Policy" (2020) 12(2) Policy & Internet 225 DOI: 10.1002/poi3.184 accessed 11 August 2018.

56: See generally Anne Helmond, "The Platformization of the Web: Making Web Data Platform Ready" (2015) 1(2) Social Media + Society 205630511560308 DOI: 10.1177/2056305115603080; Tania Bucher, "Objects of Intense Feeling: The Case of the Twitter API" (2013) 3 Computational Culture: A Journal of Software Studies http://computationalculture.net/objects-of-intense-feeling-the-case-of-the-twitter-api/; Axel Bruns, "After the 'APIcalypse': Social Media Platforms and Their Fight against Critical Scholarly Research" (2019) 22(11) Information, Communication & Society 1544 DOI: 10.1080/1369118X.2019.1637447.

57: Connie Bruck, *Master of the Game: Steve Ross and the Creation of Time Warner* (Penguin 1994) 180.

58: Zak Penn (director), Atari: Game Over (Fuel Entertainment USA, GRAiNEY Pictures, Lightbox 31 March 2015); Alex Godfrey, "'A Golden Shining Moment': The True Story behind Atari's ET, the Worst Video Game Ever" (*The Guardian*,

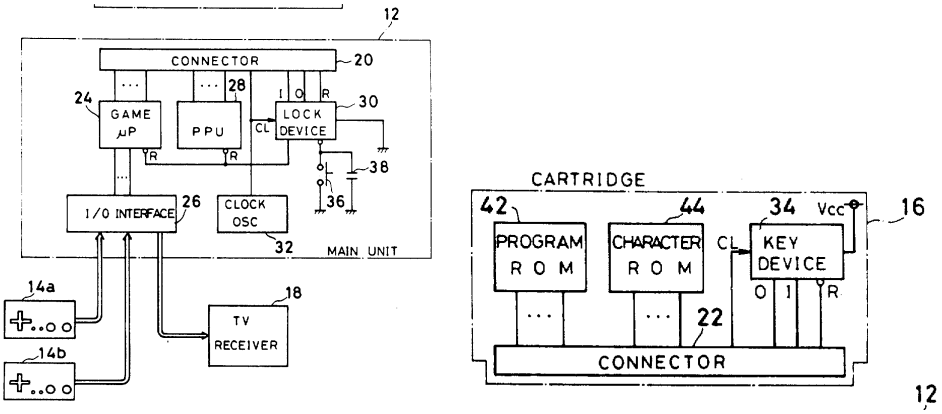

Figure 3.5: Diagram of the functioning of the 10NES chip system from Nintendo's 1985 patent. (Katsuya Nakagawa "System for Determining Authenticity of an External Memory Used in an Information Processing Apparatus" (US pat 4799635A, Nintendo Co Ltd 24 January 1989) https://patents.google.com/patent/US4799635A/en accessed 22 April 2025.) Note the lock and key chips present in the console and cartridge.

bling enthusiasm for home consoles spooked the market, leading to a 1983 crash. Atari was sold the next year, starting the rapid decline of what was briefly the fastest-growing company in U.S. history.[59]

Nintendo's response was to remodel the entire gaming hardware landscape forever to give itself the ability to vet and prevent potential E.T.-like flops. It not only changed the industry — it arguably changed the history of computing.[60] Its Nintendo Entertainment System (NES) would lock up and refuse to play game cartridges that did not have a chip in them with a corresponding key (Figure 3.5). While some previous application developers had required a piece of hardware, such as a dongle or keycard, to be present to play, Nintendo appears to have been the first to deploy a platform-wide authentication requirement designed from the outset to be difficult to bypass.[61] Nintendo combined this technoregulatory mechanism with legal protection by weaponizing copyright law: it sued those, like Atari, that sought to circumvent the NES chip's protection to sell cartridges of their own without agreeing to Nintendo's extremely stringent, arguably anti-competitive licensing conditions.[62]

This was a seminal moment: a computer manufacturer (here, making video game consoles)

30 January 2015) https://www.theguardian.com/film/2015/jan/30/a-golden-shining-moment-the-true-story-behind-et-the-worst-video-game-ever accessed 22 April 2025.

59: Mirko Ernkvist, "Down Many Times, but Still Playing the Game : Creative Destruction and Industry Crashes in the Early Video Game Industry 1971-1986" in Karl Gratzer and Dieter Stiefel (eds), *History of Insolvency and Bankruptcy from an International Perspective* (Södertörns högskola 2008).

60: Marc Steinberg, *The Platform Economy: How Japan Transformed the Consumer Internet* (University of Minnesota press 2019) (arguing Apple and Google inherit their platform strategies from Nintendo).

61: A small number of minor platforms in the 1980s are reported to have very simple checks for a copyright string in a cartridge — notably the ColecoVision console — but these were not cryptographic and trivially circumvented. Expensive software in the 1980s also occasionally came with hardware 'dongles' that required a key, or a critical part of the code, preventing the software from being run without them. See generally Ross Anderson, *Security Engineering: A Guide to Building Dependable Distributed Systems* (John Wiley & Sons, Incorporated 2020) ch 24.

62: Casey O'Donnell, "The Nintendo Entertainment System and the 10NES Chip: Carving the Video Game Industry in Silicon" (2011) 6(1) Games and Culture 83 DOI: 10.1177/1555412010377319; see the litigation in *Atari Games Corp. v. Nintendo of America Inc.*, 975 F.2d 832 (Fed. Cir. 1992).

mixed law and technoregulation to try to control who could run what software on a device that they had sold to a user. This gambit only grew in complexity after its introduction by Nintendo, and today, the most advanced practitioner of this technique is probably the multinational giant Apple. Exploring Apple's practices — some of the most restrictive forms of technoregulation applied to general-purpose computing to date — brings us closer to today's AI context.

SMARTPHONE STRATEGIES

The smartphone is today's most ubiquitous personal computer. Although these devices are hugely powerful general-purpose computing machines in theory, in practice there are typically major limits on what they can do. From the mid-2000s onward, a confluence of technoregulation, platform governance and politics has profoundly shaped the bounds of what kinds of code can, or cannot, run on mobile devices.

Apple provides perhaps the most instructive example. When the iPhone first launched in 2007, users could not install any custom software on it. In some ways, this was typical of the time — Nintendo Game Boys, digital cameras, microwaves, calculators and other compact computers all severely limited users' ability to customize them. Yet none of these devices sought to replace the huge range of personal and professional computing functions that the iPhone intended to. Apple solidified its plan to create a general-purpose computing replacement as it launched the App Store the following year. "There's an app for that," promised the firm, referring to, it seemed, practically everything.

But in fact Apple wished that there would *not* be an app for *some* things, and that meant that it needed to create a mechanism to keep unwanted software off iPhones. As a result, Apple devices began to develop the most sophisticated array of what Julie Cohen describes as "architectures of control" of any consumer computing system. Through this architecture, Apple has a decisive say over what software can and cannot run on much of what it sells.[63] This goes much deeper than the App Store, which is just the visible surface of a sprawling, subterranean walled garden. Its branches reach up into Apple's servers and its roots down into the core of its mobile devices. It manifests impressive power both to restrict users' access to the full computational potential of the devices in their pockets, and to limit businesses that wish to use certain functionality, particularly sensors, for their own commercial ends.

Apple's control extends down to the first bit of code, laid down during chip fabrication, that runs when one of its devices starts up — the boot ROM. Embedded in the boot ROM is Apple's *public key*, an openly known piece of data that allows the holder to verify digital signatures made with its corresponding *private key*, which as the name indicates, is data held secretly by the entity that wants to issue a signature. Each piece of software that loads in sequence at this point checks that the signature of the next piece corresponds to Apple's public key before allowing it to continue, creating a "chain of trust" (Figure 3.6).[64]

63: Julie E Cohen, *Configuring the Networked Self: Law, Code, and the Play of Everyday Practice* (Yale University Press 2012) http://juliecohen.com/configuring-the-networked-self (on 'architectures of control,' analogously, in pervasively distributed copyright enforcement).

64: Apple, "Boot Process for iPhone and iPad Devices" (*Apple Support*, 18 February 2021) https://support.apple.com/en-gb/guide/security/secb3000f149/web accessed 23 April 2025.

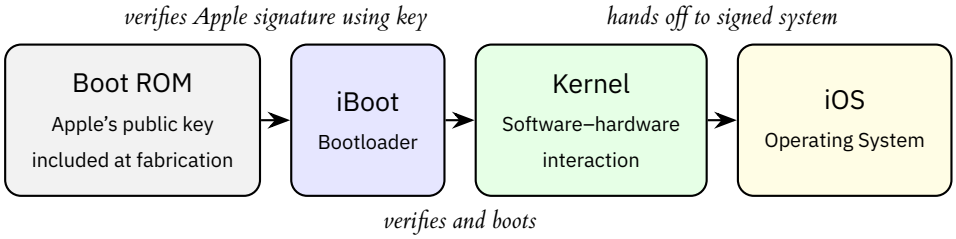

Figure 3.6: What happens when you turn an iPhone on to make sure only Apple-approved software can run — the 'secure boot chain.' Diagram: authors.

The upshot is that if Apple uses its operating system to limit the ways in which people can use software and hardware, users cannot simply switch operating systems to one that permits what they want to do. The device would refuse to load it.[65] Other checks also exist; Apple's *secure enclaves* are separate processors designed to be extremely tamper-resistant that can provide cryptographic proofs that only authorized software is running on the device.

Once the operating system loads, it shapes and defines the way any software can interact with the hardware of the device itself. Operating systems provide computational abstractions that maintain consistency across hardware. 'Files' are one such abstraction — in practice, hard drives store the data belonging to a single 'file' scattered all over the place. Files are just one of many useful lenses an operating system gives us so that we can manipulate things to easily achieve our tasks. Furthermore, abstractions facilitate important forms of interoperability. A spinning hard disk drive and a totally immobile solid state drive store files in physically very different ways, but the operating system helpfully hides all that complexity from users and software developers.

Having defined these abstractions for us, the operating system also determines the ways we interact with them. Opening a file, renaming a file, moving a file: these are all human verbs the operating system has filled with computational meaning. These pre-defined commands are another instance of a concept we saw in the previous chapter — application programming interfaces, or APIs.

On iOS, Apple does not allow all software developers to use all these API 'action verbs' or commands. Each app is installed within a constrained 'sandbox,' and every time it tries to interact with the broader operating system through an API, such as to open a file, a deeper software layer — the phone's kernel — intercepts that action and checks if the app has permission to do so. Apple calls this permission an 'entitlement.' When developers write apps, they have to declare to Apple the complete list of all the 'entitlements' they wish their app to have.

Some entitlements will never be approved without specific permission sought in advance from Apple — such as the entitlement to use the USB port to interface with custom devices; to connect to a car; or, in the time of Covid-19, to signal to others using Bluetooth while the app is not open on a user's screen, for contact tracing purposes. Such permission often comes with further conditions and policies that must be complied with, and may require further contrac-

65: Occasionally someone finds a way to do this by exploiting a bug or glitch, but these are rarer and rarer as Apple's approach has increased in sophistication.

tual agreements. Other entitlements are more widely permitted, but the operating system still remains in control, often stepping in, popping up menus to confirm that you give your consent for an app to look at your contacts or access your device's location. The firm has in the past decade developed increasingly sophisticated measures to detect and block apps accessing entitlements without permission, for instance acquiring forensic analysis firm SourceDNA in 2016 to improve its ability to sniff out disguised, unauthorized APIs in code submitted to the App Store.[66]

CUPERTINO'S RULES

Petitioning Apple, the sovereign of iOS, to grant these API permissions is a fickle business. Just like feudal kings of old, the firm often makes up rules on the fly. Nowhere is this more evident than in relation to an unspecified range of secret entitlements, sometimes called 'private APIs.'[67] The App Review Guidelines explicitly forbid developers from making use of private APIs with no stated process for exceptions,[68] yet Apple has a history of arbitrarily granting significant privileges to favored firms.

Apple grants undocumented private entitlements to its favored partners through completely nontransparent processes. We largely know about these due to a combination of forensic deduction, internal leaks and court disclosures. For instance, from the earliest days of the App Store in 2008, Google Maps was secretly and exclusively 'entitled' to use the gyrometer to trigger a voice search when you placed the phone to your ear.[69] Uber was permitted secret screen recording entitlements granted to no one else.[70] During the Covid-19 pandemic, Zoom was granted a huge competitive advantage with an exclusive entitlement allowing Zoom users to multitask on iOS devices while still appearing on camera during a call.[71]

Gatekept app stores have long struggled with a key way that malicious AI models can be distributed on iOS devices today — as downloads within an app. Consider AUTOMATIC1111, the open-ended tool used in both *bona fide* artistic communities and child abuse rings, described earlier in this chapter. AUTOMATIC1111 is a tool that runs on people's devices, accessed through their browsers, to allow individuals to generate media entirely on their devices. Similar tools, such as the app 'Draw Things,' exist in Apple's App Store. Given their specialized machine learning hardware, wide availability and ease of use, iOS devices are well-placed to run similar tools. But this is not a wholly new moderation challenge for Apple, which has generated controversy by removing many apps designed to help kids learn to code on the ground that their shell environments are prohibited forms of software-within-software.[72]

66: Evan Boehs, "App Store Levels Playing Field and Supports Developers" (26 January 2024) https://boehs.org/node/private-apis accessed 8 November 2024.

67: Apple calls these 'non-public APIs.'

68: Apple, "App Review Guidelines" (*Apple Developer*, 13 September 2024) https://developer.apple.com/app-store/review/guidelines/ accessed 4 November 2024, para 2.5.

69: John Gruber, "Google Mobile Uses Private iPhone APIs" (*Daring Fireball*, 19 November 2008) https://daringfireball.net/2008/11/google_mobile_uses_private_iphone_apis accessed 9 November 2024.

70: Boehs (n 66).

71: Jeremy Provost, "iPad Camera Multitasking" (*Think Tap Work*, 27 April 2021) https://blog.thinktapwork.com/post/6496307200846397744/ipad-camera-multitasking accessed 9 November 2024.

72: Brian X Chen, *Always on: How the iPhone Unlocked the Anything-Anytime-Anywhere Future - And Locked Us In* (Da Capo Press 2012); Brian X Chen, "Apple Rejects Kid-Friendly Programming App" (*WIRED*, 20 April 2010) https://www.wired.com/

Cupertino initially established strict, blanket rules prohibiting apps that "download code in any way or form," or that "install or launch other executable code,"[73] fearing a competitor (at the time, Adobe with its *Flash* system) nesting lots of little apps within a single app to create its own rival to the lucrative App Store.[74] Almost a decade later, Apple did change its rules slightly to permit software-within-software for limited use cases such as education, although developers still fear arbitrary, flip-flopping interpretations of the rules.[75] This fear seems justified, as Apple has gone back and forth on allowing and removing Linux emulators that let users run mini computers on their own devices and install software on them.[76]

Oddly enough, Apple does, as of early 2026, seemingly allow many apps that permit users to download any local machine learning model they want onto their devices. Apple has even given this its blessing, promoting software frameworks which allow developers to "quickly and easily update your models without updating the app itself" (which would trigger an app review process).[77] At the time of writing, the App Store has plenty of approved apps, including 'Draw Things,' which allow custom image model and LoRA installations from platforms like Hugging Face or Civitai.

Apple staff might currently see this loophole as justifiable, but that will not make it any easier for them to deal with the inevitable consequences. While Cupertino might forbid downloading executable code onto iOS, AI models are not really 'code' in the traditional sense.[78] Instead, they are more like mathematical data, consisting of weights and parameters interpreted through the software in the approved app. But they're a weird type of data — data which provides entirely new functionality; a type of data which the interpreter, unlike, for example, a Web browser interpreting JavaScript or HTML, cannot easily reason about, limit or control.

Instead, Apple might lean on its content moderation rules, which oblige apps that enable interaction with user-created content to maintain certain moderation functions.[79] Doing this

2010/04/apple-scratch-app/ accessed 3 November 2024.

73: Apple, "App Store Review Guidelines" (*Apple Developer [via Internet Archive]*, 3 September 2014) https://web.archive.org/web/20140903022336/https://developer.apple.com/app-store/review/guidelines/ accessed 3 November 2024, paras 2.7, 2.8.

74: See John Gruber, "App Store Rejection of the Week, Runner-Up: Scratch" (*Daring Fireball*, 16 April 2010) https://daringfireball.net/linked/2010/04/16/scratch accessed 3 November 2024, also claiming that Apple wanted to retain control in order to move away from Intel chips to ARM architectures without breaking everything. Today, Apple both has a specific policy for 'Mini Apps' after a long, tense stand-off with mini-app tycoon Tencent, maker of WeChat (see Pei Li, "Apple and Tencent Agree to 15% Fee on WeChat Mini Game Purchases" (*Bloomberg*, 13 November 2025) https://www.bloomberg.com/news/articles/2025-11-13/apple-and-tencent-agree-to-15-fee-on-wechat-mini-game-purchases accessed 21 December 2025), and is obliged, at least in the EU, to permit third-party app stores (see Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) [2022] OJ L265/1 (DMA) art 6(4)).

75: Thomas Claburn, "Apple Appears to Relax Ban on Apps Fetching, Running Extra Code – Remains Aloof as Always" (*The Register*, 7 June 2017) https://www.theregister.com/2017/06/07/apple_relaxes_developer_rules/ accessed 3 November 2024.

76: Theodore Dubois, Saagar Jha, and Martin Persson, "About iSH's Pending Removal from the App Store" (*iSH Blog*, 8 November 2020) https://ish.app accessed 3 November 2024; a-Shell [@a_Shell_iOS], "FYI, Apple Sent a-Shell a Similar Notice of Termination a Few Days Ago[...]" (*Twitter*, 8 November 2020) https://x.com/a_Shell_iOS/status/1325526061099196416 accessed 3 November 2024.

77: Anil Katti, "Use Model Deployment and Security with Core ML" (*Apple Developer*, 2020) https://developer.apple.com/videos/play/wwdc2020/10152/ accessed 23 November 2025.

78: Apple, "App Review Guidelines" (n 68) 2.5.2.

79: This rule has been used to exert pressure on 'alt-tech' platforms claiming to offer 'censorship-free' experiences, such as the far-right oriented Gab and Parler. 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.

might wipe the firm's hands clean of responsibility for the specific models being deployed by users, but would just kick the can of responsibility onto the app developer — a very difficult moderation challenge for actors hosting models, as we will show in Chapter 5.

ENTITLED AI GOVERNANCE

Apple has already begun to use its entitlements in ways that connect to AI governance, although not in ways explicitly intended to prevent criminal misuse of certain generative systems. The firm's 'Apple Intelligence' on-device models have a privileged entitlement providing API access to a retrieval-augmented generation (RAG) database, allowing them to access private data 'semantically indexed' across apps and services and stored on device (see Chapter 4 for more).[80] This data allows the integrated AI system to access information from third-party apps which have made certain records — such as a flight booking, or a recipe — along with certain verbs which can act upon a record (e.g. *amend*, *order ingredients for*), legible to the device.[81]

This particular restriction isn't well-suited to controlling illegal use of AI systems that have concerned us in this chapter, as their malicious capabilities do not pivotally depend on access to on-device, third-party app content. It does, however, point to the possibilities of a handset provider using the entitlement system as a governance chokepoint.

The main chokepoint Apple could use to control models is access to its on-device GPUs. Local models are rendered ineffective and inefficient — and might not have the resources to run at all — without access to the specialized hardware designed for machine learning inference and now packaged with the latest mobile devices. Apple calls the software framework that connects apps to its GPU 'Metal': it is the exclusive point of entry for apps seeking to use this ML-optimized hardware.[82] Metal is not currently governed by a regulated 'entitlement.' Any app that is running can use it without restriction. Yet given the structure of Apple's operating system, it would be trivial (albeit disruptive for existing developers) for Apple to introduce an entitlement to use the technology.[83]

Metal is not currently used as a gateway for governance. Would it even be possible in its current form? While this discussion is necessarily somewhat speculative, the first step is to understand the kinds of computational abstraction that Metal uses to connect developers with the GPUs in question. A low-level abstraction would allow developers to direct the hardware in detailed, specific ways. A high-level abstraction might, for example, see the user upload a model into the operating system that takes care of the query for the app. This parallels our discussion in Chapter 2 on types of cloud services; low-level abstractions are much more like

80: Apple, "Keynote - WWDC24" (*Apple Developer*, 9 September 2024) https://developer.apple.com/videos/play/wwdc2024/101/ accessed 9 November 2024.

81: These are called 'App Intents,' and the technology behind this kind of information is called a 'semantic graph.' At the time of writing, only the proprietary Apple Intelligence system has access to this graph and the legibility it provides. EU law may change this in the future if the enforcer gets their act together, as it requires both virtual assistants and operating systems to open up to other providers, such as alternative assistants. See DMA, art 6(7). As it stands, Apple states that developers can submit a 'request' for 'effective interoperability' but that it will not build this in by default.

82: The other methods, OpenCL and OpenGL, were deprecated in the late 2010s.

83: Despite Apple being often portrayed as an AI laggard, its hardware (Mac Mini, Mac Studio) is extremely popular and economical for AI inference compared to rivals like Nvidia.

unmanaged compute, while high-level abstractions are much more like managed or pooled compute.

Metal provides a low-level abstraction to developers, allowing them to do all but the most arcane operations with the hardware. This kind of abstraction makes governance difficult, because the operating system cannot see enough of what is going on to work out whether a model is being used in a permissible or impermissible way. The OS cannot see if a model known to be abusive has been downloaded within an app either: within the current Metal framework, the OS does not see an instruction to generate an illegal image; it instead sees a mathematical instruction given to it by the app as part of a many-stage process to query a model — for instance, 'multiply this huge matrix by this other huge matrix.' There is therefore little to no ability for Metal to reason about the intent or consequence of this calculation, and even if it could, the hardware processes involve literally trillions of operations per second. Attempting to reconstruct what all of these calculations are actually doing in concert would slow down the hardware to an unusable degree.

To make things even more difficult, generative models feature a layer called a tokenizer, which converts human-interpretable concepts, like words or tiny bits of an image, into numbers that are used in a model. Without the codebook to convert it back, even if the operating system could analyze the operations whizzing through it, it would not be able to connect them to meaningful concepts. Like the 'Wrens,' the thousands of British women working at Bletchley Park on top-secret codebreaking activities during World War II, the operating system is left doing critically important and difficult mathematical and computational work while being able to piece together surprisingly little about its origin or purpose.[84]

Restricting machine learning to only high-level APIs would allow more options for governance, giving the operating system a much greater role and visibility into the models and queries themselves. Apple does have a high-level set of APIs that operate in this way, designed for speed and energy efficiency: Core ML. In this framework, developers convert models they want to use into a special Apple format, and then pass the resulting bundle to the operating system as a whole. The OS then determines the best way to compute the results the user wants with it using the hardware available to it, including using the 'Neural Engine,' a specialized chip on iOS devices that cannot be used in any other manner. Apple does not currently force people to use this API, and if it did, it would have to do so carefully so as not to break apps using the GPU via Metal for non-AI functions, such as game graphics.

But let's posit that Apple (or other mobile device manufacturers) were to create this new chokepoint, forcing AI model querying into a high-level API with the sole aim of allowing the operating system to reason about them. Would this be effective? Even though the OS could now identify and probe the AI model, no analytic technique currently exists to 'open up' models and easily understand whether they have, or do not have, malicious capabilities.[85] Instead, researchers usually test models by benchmarking their performance on datasets designed to elicit dangerous responses, which requires many, many queries.[86] These evaluations,

84: Mar Hicks, "War Machines: Women's Computing Work and the Underpinnings of the Data-Driven State, 1930–1946" in *Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing* (MIT Press 2017).

85: Sharkey and others (n 50).

86: See e.g. Zhexin Zhang and others, "SafetyBench: Evaluating the Safety of Large Language Models" in Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume*

or 'evals,' require extensive and often expensive computing power, and are taxing even for a university lab, let alone for a pocket-friendly consumer device to do on its own. Perhaps in the future, mobile devices will be able to do this, and maybe even in a clever distributed manner to save resources, but that remains to be seen. Furthermore, this kind of governance development would pose complex legal questions. Do users get a legal safe harbor when their devices automatically and routinely try to generate illegal content such as child abuse material that may attract some of the strictest forms of liability around?

A potential device-GPU chokepoint creates two other governance options, however: model notarization and the creation of a bottleneck for our next type of substantive on-device governance, operating system surveillance.

MODEL NOTARIZATION

A high-level API for AI querying would enable the development of a system where models could be *notarized* just like apps are.

Software notarization is a special form of digital certification typically accompanied by a form of technical enforcement. It involves a developer submitting new software they wish to distribute to a signer, who checks the software against a set of policies they wish to enforce. If they think these policies are met, the signer attaches a verifiable cryptographic certificate to the software. When a user downloads it, that user's operating system looks at the software, the signature and verified public keys of the signer, checking whether the signature is valid and whether it corresponds to the downloaded code. If it does, it can run; if it does not, it may be blocked. Apple unironically calls this system "Gatekeeper" — the same term that the Digital Markets Act uses to refer to actors with market power in need of regulation.[87] On macOS and more recently on Android, running software blocked due to the lack of notarization can be manually overridden by the user, although the system is designed so that the less-savvy might not know how to do so.[88] On iOS, there is no way to bypass the blockade.

Model notarization poses various conceptual and technical difficulties. First, there is a policy question. Should the fingerprints of approved models be notarized as good and allowed, or should bad, blocked models have their fingerprints logged as bad and disallowed? A model approval process has more impact on freedom to compute, adding moderation costs that someone has to bear and restrictions that will affect consumers, businesses, hobbyists and researchers. A blocking approach has much more limited impact on generativity, but is much more fragile and easily circumvented.

This is compounded by a key difference between apps and models. Models are designed to be changed through processes like fine-tuning and LoRAs. For approved models, would fine-tuning make an approved model different, and therefore unapproved? The reason iOS devices have frameworks like Core ML and the 'neural engine' is, in large part, to enable computationally efficient fine-tuning and model personalization. For blocked models, would fine-tuning (or even trivial transformations, such as common compression or the reshuffling

1: Long Papers) (ACL 2024, Association for Computational Linguistics August 2024) DOI: 10.18653/v1/2024.acl-long.830.

87: Apple, "Notarizing macOS Software before Distribution" (*Apple Developer Documentation*) https://developer.apple.com/documentation/security/notarizing-macos-software-before-distribution accessed 3 November 2024.

88: The Android situation is live, changing and controversial, see https://keepandroidopen.org/.

of neurons within a layer) unblock a model that has exactly the same capability for misuse? Illegal image blocklists commonly used across the platform ecosystem have had to deal with these kinds of issues too. Platforms rely on vast datasets holding statistical representations of terrorist material or child abuse content, such as those maintained by the Global Internet Forum to Counter Terrorism or the National Center for Missing and Exploited Children, queried rapidly behind the scenes each time you upload something to Instagram, Facebook or a similar service. These systems use techniques that attempt to detect matches even when content has been modified (cropped, filtered, captioned, etc.).[89] Unfortunately, while we can clearly reason about what makes one collection of pixels visually similar or different from another, when it comes to AI models, we lack the analytic tools to efficiently and reliably understand the consequences of a specific model modification by observing its weights. There are many ways in which someone may have modified a model to undo safety protections while otherwise maintaining its functionality. If operating systems were to block any model that has been modified, they would also block many legitimate and useful local deployments that keep the guardrails intact.

Cryptographic hashing is the technique best known from the software notarization context, but it is not the only way to check whether a model has been changed or not. Some approaches to 'model fingerprinting' have been proposed. These either try to capture some difficult-to-change mathematical element about the model weights, or some tell-tale way a model responds to prompts. Some methods are designed to be robust to changes such as fine-tuning, and to detect what the base model was for the purposes of combating the stealing or unlicensed reuse of models subject to intellectual property protection.[90] Others are more fragile, designed to detect whether a model has or has not been tampered with.[91] These techniques increase the toolkit for AI notarization, but none of them solve the core problem: which modifications to a previously tested model actually make, or don't make, a difference to how it performs on subsequent tests?

Many policy trade-offs will thus hinge on the nature of the tests being carried out. Could such tests be automated? Can they be run on device, or would they require models (and fine-tuned versions) to be uploaded to expensive cloud servers, potentially to be 'red teamed' themselves by human reviewers or LLMs designed to coax out harmful outputs? Who would pay for this? Are such automated tests even reliable enough for the job, given that systems could be fine-tuned in a covert manner to hide specific malicious capabilities behind ostensibly innocuous prompts?

89: Hany Farid, "An Overview of Perceptual Hashing" (2021) 1(1) Journal of Online Trust and Safety DOI: 10.54501/jots.v1i1.24; Robert Gorwa, Reuben Binns, and Christian Katzenbach, "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance" (2020) 7(1) Big Data & Society 2053951719897945 DOI: 10/ggsfrk .

90: See e.g. Jiashu Xu and others, "Instructional Fingerprinting of Large Language Models" in Kevin Duh, Helena Gomez, and Steven Bethard (eds), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (NAACL-HLT 2024, ACL 2024) DOI: 10.18653/v1/2024.naacl-long.180.

91: Zecheng He, Tianwei Zhang, and Ruby Lee, "Sensitive-Sample Fingerprinting of Deep Neural Networks" in (Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019) DOI: 10.1109/CVPR.2019.00486.

COMPUTING FREEDOMS, COMMERCIAL GAINS

Given these difficult trade-offs, the wide adoption of a model notarization regime would have a significant impact on the public's freedom to compute. We have seen this movie before: the last two decades' global battles over copyright. Digital consumer technologies gave individuals the means to copy and distribute materials (such as movies, games or music) at scale extremely cheaply. Rights-holders have long been trying to restore the friction that these devices took away. Their initial strategy was to do deals with device and storage media manufacturers to have them put in mechanisms that limit individuals' ability to copy ('rip') the material freely — commonly known as 'digital rights management' technologies. One well-known example is the encryption system that an association of major Hollywood studios had developed for DVDs in the 1990s, keeping the keys private and only sharing them with DVD-player manufacturers after they signed licensing and non-disclosure agreements. As with Nintendo a decade earlier, the copyright lobby worked simultaneously to enshrine civil or even criminal penalties in national and international law to place liability on anyone who tried to tamper with these often flaky technical protections, even if they were doing it for a legally permitted purpose, such as parody or news reporting.[92]

As narrow-purpose VHS players and TV recording boxes gave way to open-ended home computers, a more ambitious version of the same tactic was tried, branded 'trusted computing.'[93] In the early 2000s, Microsoft attempted to seize control of end-user computation with a trusted computing proposal for Windows called *Palladium*. In essence, Palladium sought to ensure that a third party with which a device is interacting can verify that the program in question has not been modified by the user — such as to allow an encrypted video to be copied into a new format after the software had decrypted it. While industry consortia sold it as an advance in security, critics branded it 'treacherous' or 'controlled' computing.[94] Influential Cambridge computer security scholar Ross Anderson argued that the media industry would use it to sell you DVDs you can't copy, downloads you can't swap and "CDs that you'll only be able to play three times, or only on your birthday."[95] Digital rights advocates have remained firmly opposed to expansive digital rights management, arguing that fundamental restrictions on computing are an unacceptable constraint on civil liberties. Relatedly, those critical of the maximalist status quo in copyright, ranging from scholars to Pirate Party Members of the European Parliament, argue that efforts to make it illegal to even try to get around technoreg-

92: See generally Julie E Cohen, "WIPO Copyright Treaty Implementation in the United States: Will Fair Use Survive?" (1999) 21 European Intellectual Property Review 236; Pamela Samuelson, "Intellectual Property and the Digital Economy: Why the Anti-Circumvention Regulations Need to Be Revised" (1999) 14(2) Berkeley Technology Law Journal 519. For specific provisions, see World Intellectual Property Organization Copyright Treaty (adopted 20 December 1996, entered into force 6 March 2002) 2186 UNTS 121, art 11; 17 U.S.C. §1201 (for U.S. anti-circumvention regulations); Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L167/10, art 6 (for the EU equivalents). The DVD Copy Control Association's 'Content Scramble System' algorithm was 'jailbroken' by hobbyists three years after release, but later DRM systems built into e.g. Blu-Ray are more sophisticated and feature substantially stronger encryption. For more, see https://en.wikipedia.org/wiki/Content_Scramble_System.

93: Cohen, *Configuring the Networked Self: Law, Code, and the Play of Everyday Practice* (n 63).

94: Richard M Stallman, *Free Software, Free Society* (Free Software Foundation 2015) ch 37; Ross Anderson, "Cryptography and Competition Policy: Issues with 'Trusted Computing'" in *Proceedings of the Twenty-Second Annual Symposium on Principles of Distributed Computing* (PODC '03, Association for Computing Machinery 2003) DOI: 10.1145/872035.872036.

95: Anderson, "Cryptography and Competition Policy" (n 94).

ulation unsettle the careful balance previously struck by copyright law between proprietary control and public and societal interest.[96]

Microsoft's Palladium project never actually materialized, but controlled computing has not gone away. Underpinned by the secure hardware and boot chains described above, the copyright aspects mostly moved into the Web browser as many people abandoned purchased files for subscription streaming services. Try to take a screen recording of Netflix on your computer or mobile device, and you'll likely find it comes up blank — a result of a complex stack of technologies that now connect browsers to graphics chips.[97]

This raises a crucial point. If regulators were to introduce model notarization into our devices for the purposes of preventing socially harmful and even dangerous misuse of those models, this intervention would conjure an entire infrastructure that would surely be cynically repurposed for private profit. The creators of downloadable models, or the platforms they are provided through, could then surveil and control the use of these models to widen their own profit margins. Software notarization is not a value-free technology — quite the opposite. Consider its role in allowing Apple to charge exorbitant fees to developers for any in-app purchases — often 30% of the developer's take, amounting to tens of billions of dollars for the firm.[98] Consider also digital rights management's role in creating a market for 'downloadable' videos, books and music that users cannot play if a monthly subscription is not paid. Both technologies were ostensibly developed in the name of public interest — Apple claiming to protect consumer interests by creating cancellation and other rights, and media platforms claiming to fight piracy.[99]

OPERATING SYSTEM SURVEILLANCE

Forcing all actors to use specific device APIs for model querying could also generate mandatory bottlenecks where operating system–led scanning could occur. If the operating system has to see the content of all queries and all outputs that come in and out of the models it has authorized to run, then it could also add a layer of automated scanning or transformation, much like major cloud providers already do. These technologies (and their limitations) are dealt with in detail in Chapter 2, where we looked at filtering and transformation for both content moderation and for provenance purposes, such as through watermarking. Such functionality attached to an API is sometimes called a form of 'middleware' — software sitting in between an app and the hardware.[100] If there is no way to access the hardware on the device

96: Dan L Burk and Julie E Cohen, "Fair Use Infrastructure for Rights Management Systems" (2001) 15 Harvard Journal of Law & Technology 41.

97: We don't have the space to discuss these at length here, but for the interested, these are called FairPlay (Apple), Widevine (Google) and PlayReady (Microsoft).

98: Apple's App Store was estimated by consultancy Sensor Tower to bring in $43.7 billion in the first half of 2022. Double that, and take 30%, and you get $26.2 billion as a ballpark estimate of Apple's own revenue from this source. See Stephanie Chan, "Global App Spending Reached $65 Billion in the First Half of 2022, Up About 1% Year-Over-Year" (*Sensor Tower*, June 2022) https://sensortower.com/blog/app-revenue-and-downloads-1h-2022 accessed 3 November 2024. Apple has since halved the commission for smaller developers, but this does little to change the magnitude of revenue brought in.

99: On the former, see Order on Motion for Attorney Fees, *Epic Games, Inc. v. Apple Inc.*, No. 4:20-cv-05640 (N.D. Cal. Apr. 30, 2025), ECF No. 1508, pp. 35–8.

100: Harold Abelson and others, "Bugs in Our Pockets: The Risks of Client-Side Scanning" (2024) 10(1) Journal of Cybersecurity tyad020 DOI: 10.1093/cybsec/tyad020, 11.

for the purpose of querying AI systems other than this high-level API, all inputs and outputs to on-device AI must therefore pass through the governing middleware — a sort of digital control room with deep visibility into everything going on across the system.

With such visibility into activities happening on a computing device, even at modern speed and scale, one can scan. Scanning amounts to either a process of matching or a process of prediction.[101] Matching techniques have been at the center of controversies around 'client-side scanning' in the context of end-to-end encrypted messenger services. In a world where we all have easy access to cheap, relatively robust and surveillance-resistant channels for communicating privately with one another — freely available apps like WhatsApp and Signal — the dominant law enforcement status quo has shifted away from Snowden-era mass surveillance and toward targeted hacking of devices belonging to a smaller number of suspicious individuals.[102] (It doesn't matter if your Signal messages are encrypted in transit between you and Signal if law enforcement, or a state intelligence agency, has placed malware on your device that lets them access everything that you can.) Nonetheless, security-focused actors and interest groups continue to advocate for laws that would mandate bulk scanning for certain forms of illegal activity. At the forefront of these debates has been the specter of child sexual abuse imagery; new regulation in the UK and draft regulation in the EU mandate that devices report, block and refuse to send images that trigger a match in an on-device database of perceptual hashes of previously reported child abuse material.[103]

Such techniques have led to a range of criticism. First is the question of actual efficacy. This kind of scanning is trivially avoided by putting an image in an encrypted zip file that anyone can make with a few clicks on any laptop. More sophisticated techniques include a recent evasion attack reporting a 99.9% success rate, even without access to the inner workings of the detection system.[104] On-device scanning systems are typically less reliable than on-server ones in order to make them efficient and compact, and even a small error rate can entail millions of false positives at platform scale — false positives that not only overwhelm the authorities but also potentially cause significant arbitrary harm to innocent users who can have their access to services removed.[105] Another common argument is that these scanning techniques will also become mandatory in jurisdictions where political speech is widely suppressed, and thus leveraged to quietly memory-hole various media at the regime's will. Similar hash-matching tools are reportedly deployed for information control purposes in China by Tencent's WeChat.[106] Hash matching is currently used primarily to police public platforms,

101: Carey Shenkman, Dhanaraj Thakur, and Emma Llansó, "Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis" (*Center for Democracy and Technology*, 20 May 2021) https://cdt.org/insights/do-you-see-what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/ accessed 23 December 2025.

102: Joseph Cox, *Dark Wire: The Incredible True Story of the Largest Sting Operation Ever* (PublicAffairs 2024).

103: In the UK, this is based on Online Safety Act 2023, s 121; in the EU, there have been many attempts to mandate this as part of the proposed Child Sexual Abuse Regulation, although at the time of writing mandatory scanning within encrypted environments has been dropped due to ongoing controversy.

104: Shubham Jain, Ana-Maria Crețu, and Yves-Alexandre de Montjoye, "Adversarial Detection Avoidance Attacks: Evaluating the Robustness of Perceptual Hashing-Based Client-Side Scanning" in (31st USENIX Security Symposium (USENIX Security 22), 2022) DOI: 10.48550/arXiv.2106.09820.

105: Sebastian Meineck, Markus Reuter, and Andre Meister, "Geleakter Bericht: EU-Kommission nimmt hohe Fehlerquoten bei Chatkontrolle in Kauf" (*Netzpolitik*, 29 June 2022) https://netzpolitik.org/2022/geleakter-bericht-eu-kommission-nimmt-hohe-fehlerquoten-bei-chatkontrolle-in-kauf/ accessed 23 December 2025.

106: Riana Pfefferkorn, "Client-Side Scanning and Winnie-the-Pooh Redux (Plus Some Thoughts on Zoom)" (*Stanford CIS*, 12 May 2020) https://cyberlaw.stanford.edu/blog/2020/05/client-side-scanning-and-winnie-pooh-redux-plus-some-

and happens in the cloud, at the platform's computational expense; putting this infrastructure into all of our pockets would hypercharge the already very political questions at the heart of online content governance.[107] Which actors can get media uploaded to scanning tools globally or in certain jurisdictions? Since this scanning would occur at the device level, making it potentially impossible to open certain photos, files and more, this would be a powerful tool for political and social control indeed.[108] Furthermore, researchers have demonstrated that some hash-based scanning methods, designed to be robust against attempts to tweak content in order to fool known detectors of image-based sexual abuse, are themselves vulnerable to reverse engineering that can recover variants of the original images, which can serve to increase risk and perpetuate harm.[109]

That all said, these kinds of matching tools which search for already reported content are actually not very useful for combating *generative AI* misuse. The activity at the heart of this book by nature produces never-before-seen, new content that may indeed be derivative, but nonetheless is not exactly like anything else in existence. You can't use hashing to search for things that the system hasn't seen before. This means that any scanning regime would require additional AI models, probably also on the device, that would try to make predictive classifications about the content fed to them. Are these skin-toned pixels in a generated image or video correlated with nudity? Is that object in someone's hand a weapon, or a utensil? Such tools are computationally weaker and more error-prone versions of those we discussed in the cloud context in Chapter 2. As we outline there, these technologies are both prone to simplistic overblocking and vulnerable to 'jailbreak' methods that bypass the protections they intend to impart. Sitting on-device, these kinds of scanning systems are also likely to be much easier for adversaries to analyze as they try to learn how to bypass them.

As also discussed in Chapter 2, a core part of the challenge facing us is that the dangerous misuse of AI models may not be apparent from single pieces of generated content: analysis over a period of time, and in the broader context of the device's usage, will often be needed to make accurate classifications. In this vein, the UK government has expressed a desire to have operating systems universally scan content, regardless of whether it is contained within an app or elsewhere on the device, by constantly monitoring and analyzing what is displayed on users' screens.[110] Building in such a vantage point for the operating system would only make it easier for this capability to be piggybacked upon by attackers, and would expose personal devices to enormous security risks:[111] an approach for preventing misuse that would involve the rollout of tools resembling the Pegasus spyware deployed by the NSO Group. With Pegasus, an operator need only enter a phone number — the target need not open a link, or interact in any way with the payload subsequently sent to them — to exploit a phone and obtain everything from full photo access to live location data and remote microphone control. Activists have

thoughts-zoom/ accessed 23 December 2025.

107: Robert Gorwa, *The Politics of Platform Regulation: How Governments Shape Online Content Moderation* (Oxford University Press 2024).

108: Abelson and others (n 100) 11.

109: Sophie Hawkes and others, "Perceptual Hash Inversion Attacks on Image-Based Sexual Abuse Removal Tools" (2025) 23(3) IEEE Security & Privacy 64 DOI: 10.1109/MSEC.2024.3485497.

110: Chris Smyth and Anna Gross, "UK to Push for Nudity-Blocking Software on Devices to Protect Children" (*Financial Times*, 15 December 2025) https://www.ft.com/content/0ef79775-eadf-4cc9-b32c-e97b0eff816f accessed 24 December 2025.

111: Abelson and others (n 100).

been detained and journalists killed with the help of this technology.[112]

Interestingly, these developments aren't purely driven by law enforcement and national security officials. State actors aren't the only ones that wish to better monitor what people are doing with their devices: in some cases, industry is already willingly building these kinds of tools. Microsoft, for example, has tried to deeply integrate AI into its PCs and other devices, stating grandiosely that "Windows is evolving from an operating system into the canvas for AI."[113] One of the most ballyhooed new features in its recent Copilot-enabled laptops is Windows Recall, which constantly logs what is on the user's screen, in order to allow various integrated models to interpret, categorize and make inferences about user behavior. While this 'god mode' visibility into what you're doing with your laptop is being marketed as a benign productivity tool — FAQ documents feature use cases like finding "that pizza recipe you saw earlier today but you don't remember where you saw it"[114] — it has been thoroughly roasted by privacy and security professionals. Even after a year of improvements to the system, and to the way that these snapshots were handled and stored, insecurities remain. Some analysts strongly recommend that it be disabled by journalists, minoritized groups worried for their safety, people crossing borders where device checks might occur and people in potential intimate partner abuse situations.[115] The magnitude of these concerns implies that such tools should not be mandated in sensitive environments or rolled out universally.

SEGMENTED TECHNOREGULATION

To summarize: tools that limit what can be done on individuals' devices do exist, but they are intricate, challenging to deploy and may not actually be up to the task of preventing malicious AI use. Major shortcomings include their inability to effectively detect misuse in the first place, and/or their tendency to over-flag suspicious behavior, raising proportionality and legitimacy questions.

We have primarily worked with the example of Apple, and specifically its iOS platform, as a proof-of-concept of how restrictive local devices might operate. But as it stands, most consumer laptops and desktops are not nearly as constrained as iOS. This is even more true for the long tail of more open-ended hardware aimed at hobbyists, tinkerers, businesses or researchers (think for instance of Linux users, or open-source aficionados building alternative, privacy-preserving mobile OS suites like Graphene).

The policy benefit of having such protections on only *some* devices is also limited. Assuming that in an open market all devices have roughly comparable hardware, determined malicious actors will simply obtain devices with fewer restrictions. But binding legal restrictions on

112: Laurent Richard and Sandrine Rigaud, *Pegasus: How a Spy in Your Pocket Threatens the End of Privacy, Dignity, and Democracy* (Henry Holt and Company 2023).

113: Stefan Kinnestrand, "Evolving Windows: New Copilot and AI Experiences at Ignite 2025" (*Windows IT Pro Blog*, 18 November 2025) https://techcommunity.microsoft.com/blog/windows-itpro-blog/evolving-windows-new-copilot-and-ai-experiences-at-ignite-2025/4469466 accessed 7 January 2026.

114: Microsoft, "Retrace Your Steps with Recall" (7 January 2026) https://support.microsoft.com/en-us/windows/retrace-your-steps-with-recall-aa03f8a0-a78b-4b3e-b0a1-2eb8ac48701c.

115: Kevin Beaumont, "Microsoft Recall on Copilot+ PC: Testing the Security and Privacy Implications" (*DoublePulsar*, 21 April 2025) https://doublepulsar.com/microsoft-recall-on-copilot-pc-testing-the-security-and-privacy-implications-ddb296093b6c accessed 24 December 2025.

all open-ended computing devices seem like a non-starter, even if the political will (and the market power needed to force the major device manufacturers to actually implement the desired regime) existed. Interest groups pushing for deeper hardware controls might be less concerned about how their proposals affect AI research, which requires custom computing set-ups and high efficiency. However, many other businesses and parts of government also need software which is specially hardened against security threats and surveillance, including state surveillance, from lawyers needing to talk confidentially with clients to organizations managing information critical for national security. Even with safety systems as a major motivator of on-device controls, these systems have to be tested and hardened, including using tools which generate the same content they are trying to block. If *every* device were somehow subject to filtering and control, then how would these safety systems even be trained, tested and maintained?

There would also be competition policy implications to strong device-level governance — mandatory chokepoints would extend the often-criticized platform control of Apple's mobile devices to other domains and companies — as well as principled arguments relating to the freedom to compute and tinker.[116] We would be living in a vastly less generative world (again, in a Zittrainian sense) — in some cases, for better, but in many other cases, much worse.

So what kinds of governance would actually be feasible? Instead of applying restrictions on all devices, one approach might be to segment the availability of different types of computing. This would see a specialized market of less- or unrestricted devices available to authorized entities (such as researchers, governments or *bona fide* businesses with a need for them). The general market available for consumers to purchase publicly would consist of devices that feature some kinds of protections against AI misuse, restricting the configurability and generativity of this hardware.

GOING HAM

Such a policy regime is not entirely unprecedented. Consider the regulation of two-way radio equipment in the United States: a combination of tiered licensing and design-based regulation has been deployed in an attempt to provide controlled access to wireless communications while avoiding misuse and abuse. Commercial players, big or small, have to engage with complex and often quite expensive licensing conditions, and are closely watched by the regulator. Amateur ('ham') radio operators — a term originating in their apparent "ham-fisted" propensity to clog the airwaves with bad Morse code[117] — are allowed to use their own equipment on a range of amateur frequencies depending on which tier of exam they take. The frequencies most capable of misuse and global disruption are only available to those certified after passing more difficult examinations. However, alongside this barrier to entry, some of the equipment sold even to these licensed individuals is already limited by

116: Ed Felten, "The New Freedom to Tinker Movement" (*CITP Blog*, 21 March 2013) https://blog.citp.princeton.edu/2013/03/21/the-new-freedom-to-tinker-movement/ accessed 26 December 2025; Pamela Samuelson, "Freedom to Tinker" (2016) 17(2) Theoretical Inquiries in Law 562 DOI: 10.1515/til-2016-0021; D Sculley and others, "Hidden Technical Debt in Machine Learning Systems" in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2* (NIPS'15, 2015).

117: See https://en.wikipedia.org/wiki/Etymology_of_ham_radio for a fuller exposition.

design to prevent misuse; for example, since the 1990s, all new receivers sold are required to block analog mobile telephony bands for privacy reasons.[118] This regulated hobbyist tier is combined with a set of unlicensed technologies that can be used more freely by consumers because they are so heavily constrained by design, such as the 'Citizen's Band' (CB) radio, which had its heyday in the 1970s with 15–20 million users.[119] Only certified CB devices are legal to sell, and these have extremely limited power that is illegal to increase through modification. 'Family Radio Service' (FRS) devices, which, unlike CB, make use of a part of the spectrum suited for small antennas, such as those on handheld devices rather than a car or building, are even more locked down by manufacturers. Through extreme design features, such as non-removable antennas and power limits, it is almost impossible to cause significant disruption with these certified, supermarket-bought walkie-talkies.[120]

DRONING ON

A similar approach has taken shape when it comes to other open-ended technologies that are ripe for misuse. Consider the European regulation of unmanned aerial vehicles (UAVs), more commonly known as drones. The policy is simple: less dangerous drones, which weigh less than 250 grams and have a certified altitude limit of 120 meters, can be purchased and flown by anyone. Slightly heavier drones must broadcast location and ID, and integrate a location warning system called 'geo-awareness,' using pre-stored, nationally specific data to warn the pilot when it approaches or enters prohibited airspace.[121] Beyond a medium weight, drones get even more socially risky, and here the EU regulatory regime seeks to control the operator directly rather than indirectly controlling them through measures baked into the product. To fly the heaviest and most dangerous devices, European law requires users to submit to tiered levels of training and certification, including proctored written tests and supervised practical examinations.[122]

⋆

Might this approach to segmenting the market of risky and non-risky devices be effective for the governance of on-device misuse of AI systems? The answer will be heavily contingent on a range of factors.

COORDINATION CHALLENGES

The first problem is whether or not the types of safeguards we discuss could actually be adopted across the consumer or non-specialist computing device industry. As depicted earlier in this chapter (Figure 3.6), the relevant protections start at the hardware level, and thus

118: 47 USC §302a. Today mobile phone signals are both digital and encrypted and so this is less of a concern.

119: Kristen Haring, *Ham Radio's Technical Culture* (MIT Press 2006) DOI: 10.7551/mitpress/3405.001.0001 p. 155. Formally, CB *did* require a license until 1983, but it was granted to anyone who asked for a small fee and with no examination.

120: 47 CFR §95.587.

121: Commission Delegated Regulation (EU) 2019/945 of 12 March 2019 on unmanned aircraft systems and on third-country operators of unmanned aircraft systems [2019] OJ L152/1 (elegated Regulation (EU) 2019/945).

122: Commission Implementing Regulation (EU) 2019/947 of 24 May 2019 on the rules and procedures for the operation of unmanned aircraft systems [2019] OJ L152/45.

will require significant supply chain governance. A legal mandate for design-based regulation might run into challenges in some jurisdictions, for example in the United States, where it may find itself challenged as unconstitutional under the free speech clause of the First Amendment.[123] Given the indirect link to consumer protection aims, device firms will, without a regulatory mandate, have very limited incentive to add in these extensive protections themselves.

However, insecurity against malicious actors has previously been one of the main drivers of truly international technology law. In 2000, billions of dollars of damages were caused by the ILOVEYOU virus, as a twenty-three-year-old man in the Philippines devised a worm intended to steal Internet dial-up credentials from the rich and distribute them to the poor. The virus — originally his undergraduate thesis proposal that was rejected on ethical grounds — propagated across the world as email users clicked attachments which appeared to be confessional letters from friends or colleagues, promptly forwarding the virus to everyone in the unwitting victim's address book.[124] At the time, the Philippines had no laws criminalizing hacking, much to the chagrin of high-income countries in Europe and North America. The event would push the previously stalled Budapest Convention on Cybercrime over the line, which today remains one of the most important international frameworks relating to digital technologies.[125] It remains the first and only Council of Europe treaty to ever be ratified by the United States.

CHIP CONTROLS

Even if international law or the action of large markets like the European Union were to push most device manufacturers to build in local safeguards for the consumer market, the second problem pertains to the challenge of meaningfully restricting access by malicious actors to the specialist, more open-ended hardware that would still need to be available for corporate, government and research use.

Until now, we have not discussed the existing restrictions that scholars and commentators have proposed on core technologies underpinning the current generative AI gold rush, such as powerful chips for training. This is because they come from researchers in 'AI safety,' a field which overwhelmingly concerns itself with preventing AI capabilities from leading to dangerous and unhappy *accidents*, rather than malicious use by ordinary users or criminal entities.[126] When it comes to difficult questions around hardware governance, the central thrust of 'AI safety' discourse has thus almost entirely been about attempts to restrict computing power, rather than more nuanced conversations about what that power is actually used for. Restricting certain countries from obtaining GPUs does little to stop the difficult-to-prevent

123: *NetChoice, LLC v. Bonta*, 113 F.4th 1101, 1112 (9th Cir. 2024); see generally Rebecca Aviel and others, "From Gods to Google" (2025) 134 Yale Law Journal 1269 DOI: 10.2139/ssrn.4742179.

124: Geoff White, *Crime Dot Com: From Viruses to Vote Rigging, How Hacking Went Global* (Reaktion Books 2020).

125: Convention on Cybercrime (Council of Europe) (opened for signature 23 November 2001, entered into force 1 July 2004) 185 ETS; Shannon C Sprinkel, "Global Internet Regulation: The Residual Effects of the ILOVEYOU Computer Virus and the Draft Convention on Cyber-Crime" (2001) 25(3) Suffolk Transnational Law Review 491.

126: See generally Karen Hao, *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI* (Penguin Press 2025) ch 2; Dario Amodei and others, "Concrete Problems in AI Safety" (*arXiv*, 25 July 2016) DOI: 10.48550/arXiv.1606.06565 (a paper defining AI safety as accident avoidance, from the now-CEO of Anthropic).

misuse of highly capable non-frontier systems for fraud, political and social deception and sexual abuse.

Some of these forms of hardware governance mirror what we have discussed in this chapter. Instead of trying to stop users from doing specific *things* with the systems they hold, 'AI safety' research has led to policy that focuses on preventing unauthorized users from accumulating usable computing power beyond a certain threshold.[127] Other related governance methods focus on limiting the computing power that can be accumulated even if large numbers of chips are owned, such as by limiting the number of other chips they can network with at once.[128] Unlike in the doomsday scenarios that keep Effective Altruists and 'AI doomer' researchers up at night, there is no linear causal relationship between more chips and more or better fraud. Readily available top-line PCs and even mobile devices are already fit for purpose; a canny deepfakist harassing you in 2030 will probably only need one Nvidia GPU, not hundreds of thousands. Some other hardware governance mechanisms proposed by AI safety researchers focus instead on who is an authorized user of such chips. They typically either build reporting or tracking into chips, or ensure such chips do not work unless certain conditions are met. For example, chips might stop working after a certain amount of time or use unless they receive a renewed cryptographic license from the manufacturer.[129] They might also stop working if they detect they are in a certain prohibited country, or at least attempt to report the fact to their manufacturer.[130] Technology to geolocate chips, typically by seeing how long it takes to send a round-trip message to several servers in known physical locations, is already deployed by Nvidia in its optional GPU fleet management software.[131] The 'Chip Security Act,' a bipartisan bill introduced into the U.S. House of Representatives in 2025, would, if passed, mandate such systems, although it does not go so far as requiring chips to stop functioning if in prohibited jurisdictions, presumably because such a "kill switch" would create a serious

127: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L (AIA) art 51(2) (a presumption of "high-impact capabilities" where training computation exceeds 10^{25} floating point operations); President of the United States of America, "Executive Order 14110 of October 30, 2023: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (2023) 88(210) Federal Register 75191 https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence accessed 20 June 2024, para 4.2(b), placing (since repealed) reporting obligations where training exceeds the higher threshold of 10^{26} floating point operations (10^{23} for models trained primarily on biological sequence data); on why this is generally a bad idea, see generally Sara Hooker, "On the Limitations of Compute Thresholds as a Governance Strategy" (*arXiv*, 29 July 2024) DOI: 10.48550/arXiv.2407.05694.

128: Gabriel Kulp and others, "Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090" (*RAND Corporation*, Santa Monica, CA, 18 January 2024) https://www.rand.org/pubs/working_papers/WRA3056-1.html accessed 5 November 2024; Girish Sastry and others, "Computing Power and the Governance of Artificial Intelligence" (*arXiv*, 13 February 2024) DOI: 10.48550/arXiv.2402.08797; Onni Aarne, Tim Fist, and Caleb Withers, "Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing" (*Center for a New American Security*, Washington, DC, 8 January 2024) https://www.cnas.org/publications/reports/secure-governable-chips accessed 5 November 2024.

129: Kulp and others (n 128).

130: Aarne, Fist, and Withers (n 128).

131: Stephen Nellis and Michael Martina, "Nvidia builds location verification tech that could help fight chip smuggling" (*Reuters*, 10 December 2025) https://www.reuters.com/business/nvidia-builds-location-verification-tech-that-could-help-fight-chip-smuggling-2025-12-10/ accessed 24 December 2025; Nvidia, "Opt-In NVIDIA Software Enables Data Center Fleet Management" (*NVIDIA Blog*, 10 December 2025) https://blogs.nvidia.com/blog/optional-data-center-fleet-management-software/ accessed 24 December 2025.

security vulnerability for all users.[132]

This kind of 'global AI arms race' thinking is geopolitically inflected and beyond the scope of this book. There's a whole world of complex and contested tariffs, sanctions and supply chain restrictions that now are part of the political economy of semiconductor manufacturing and AI development more generally in the 2020s.[133] But the crucial point for our purposes is this: there is already ample research on different mechanisms to stop certain types of hardware from accruing into the hands of some actors. In the existing 'AI safety' literature, this hardware is generally understood as the leading-edge GPUs and specialized training chips, and the main concern is the concentration of hardware in the hands of adversarial state actors that may be unfriendly to U.S. interests, a rising China in particular. Some of these 'compute governance' techniques could indeed be used in a more humble fashion to track where devices with unrestricted and unmonitored AI-optimized chips end up, with the aim of restricting them predominantly to authorized professional users. Although black markets for all difficult-to-obtain goods have existed since time immemorial — and already exist for AI chips, as countries attempt to route around U.S. export controls[134] — the idea would be that restricting the supply of these devices would significantly reduce the number of bad actors that are able to get their hands on them. In this world, it could be less likely that your ex-partner is able to run a sophisticated deepfake operation from their laptop; perhaps only wealthy criminal gangs and syndicates are able to acquire expensive, diverted legitimate devices with that power on the black market.

HELPING HOOKS

We've explored a range of different forms of on-device governance mechanisms in this chapter thus far, bringing in occasional glimpses of regimes built for other technologies. If we put this all together, and speculate a little bit about the kinds of effective policy intervention that would be needed to responsibly shape local AI governance, the rough contours of a few potential paths forward emerge. For instance, one could imagine a situation where, like ham radios or consumer drones, 'safe-by-design' devices become available to the general public, while simultaneously, a highly regulated subset of industry would be allowed more open-ended, higher-risk hardware. Those allowed to use the most generative technologies freely might have access to open-ended chips/devices that are filled with trackers and conditions that promote accountability of use, while consumers would be faced with locked-down devices that would force them to allow some forms of scanning, blocking, watermarking or other limits on models they run.

132: Chip Security Act, H.R. 3447, 119th Cong. (2025). On security risks, see Harold Abelson and others, "Keys under Doormats" (2015) 58(10) Communications of the ACM 24 DOI: 10.1145/2814825, (in relation to backdoors generally); David Reber Jr, "No Backdoors. No Kill Switches. No Spyware." (*NVIDIA Blog*, 5 August 2025) https://blogs.nvidia.com/blog/no-backdoors-no-kill-switches-no-spyware/ accessed 24 December 2025, (expressing Nvidia's view on the undesirability of 'kill switches and built-in backdoors')

133: Chris Miller, *Chip War: The Fight for the World's Most Critical Technology* (Scribner 2022); Guillaume Beaumier and Madison Cartwright, "Cross-Network Weaponization in the Semiconductor Supply Chain" (2023) 68(1) International Studies Quarterly sqae003 DOI: 10.1093/isq/sqae003.

134: Eleanor Olcott and Zijing Wu, "Nvidia AI Chips Worth $1bn Smuggled to China after Trump Export Controls" *Financial Times* (25 July 2025) https://www.ft.com/content/6f806f6e-61c1-4b8d-9694-90d7328a7b54 accessed 8 January 2026.

In theory, if implemented, such an approach would create many potential loops that other areas of the stack could hook into to support their governance efforts. Operating systems could make model allowlists or blocklists enforceable. Watermarking systems implemented at a local level after high-level model APIs are queried could be checked later in the stack.

Yet the mechanisms needed to transform local devices from the 'unmanaged compute' category of the previous chapter (Chapter 2) into more governable APIs with clear chokepoints come with a huge generativity cost. It would mean that individuals and developers could only use hardware in ways the operating system permits, restricting potentially innovative new applications until they are authorized by infrastructure-owning firms and other gatekeepers. Such power is not wielded in an accountable manner today by companies like Apple; significant effort would be needed to create a fairer regime if this chokepoint were to become an important part of holistic, full-stack AI governance going forward.

The required extreme forms of software gatekeeping would mean that third-party operating systems would effectively be crippled. If hardware could run entirely on open-source, third-party operating systems (such as Linux or some forks of Android), then malicious users would simply buy devices and flash them. But prohibiting open OSes would be bad news for those seeking to push back through their individual choices on the power of proprietary technology firms, as well as for the growing movement fighting against planned obsolescence and e-waste by repurposing old devices, giving them a new life after their security updates end. While some, in response, might push for security obligations on open-source operating systems, and mechanisms to sign 'approved' OSes so that they could run on locked-down hardware, this kind of effort is likely to meet with opposition from both open-source advocates and the device manufacturers themselves.

Local scanning is also controversial, introducing a powerful chokepoint for off-stack actors, particularly national security and intelligence agencies, to intervene to exfiltrate data or to influence how these systems function beyond what might be considered legitimate counter-misuse efforts. Generative AI systems are already being implicated within everyday computational activities, such as making slides or editing photos; monitoring the functioning of those systems creates a powerful vulnerability. As these systems burrow deeper into our devices — and into confidential applications, such as encrypted messaging systems — subjecting them to controls, limits and abuse reporting requirements comes with risks, especially as operating system creators may keep the details of any governance arrangements opaque due to commercial sensitivity, government gag orders and wishes not to give too much information to malicious actors trying to bypass these measures.

The practical possibility of placing restrictions on all widely distributed devices, and implementing supply chain controls on all fully open devices, is made more difficult under conditions of economic globalization. Just a single manufacturer choosing to breach any governance arrangement would create a source of insecurity. The current geopolitical context of AI software and hardware development is fraught and non-cooperative, with the U.S. and China both vying for more independence from each other's sphere of technoregulatory influence. China reacted explosively to reports of location tracking and 'kill switches' in Nvidia chips, and larger governance moves will likely attract a more powerful reaction still.[135] The

135: Meaghan Tobin and Xinyun Wu, "China Summons Nvidia Over 'Backdoor Security' Risks of A.I. Chips" (*The New York*

huge growth of data center construction — reportedly accounting for nearly all of U.S. GDP growth in the first half of 2025 — creates clear possibilities for the exfiltration of open-ended hardware to criminal groups around the world.[136] (This is the exact mechanism that China has been accused of using to avoid U.S. export controls on Nvidia Blackwell chips.)[137] Even if the most powerful hardware could be secured, the threats this book discusses might materialize even with devices that were basic research grade, or long outdated, as long as they lacked the proposed software controls that the latest consumer devices would have in this kind of scenario.

Some forms of the local-layer control we have outlined above might be implemented in whole or in fragments in certain jurisdictions. But to robustly secure the *entire* global hardware supply chain seems like a pipe dream, as well as a restrictive, surveillant nightmare. A more likely scenario involves firms using local governance strategies as a way to wash their hands of potential liability and to add some friction to brand-undesirable criminal uses of *their* devices in particular. It will not, it seems to us, be feasible or desirable to build an overarching, internationally functioning regime designed to comprehensively disrupt the serious misuse of local devices by motivated, resourceful and organized actors. It is not likely to work as well as envisaged, and will therefore be unlikely to be proportionate to the consequences it brings. Other layers of the stack will have to pick up the slack.

Times, 31 July 2025) https://www.nytimes.com/2025/07/31/business/china-nvidia-h20-chips.html accessed 31 December 2025.

136: Louise Matsakis, "The AI Data Center Boom Is Warping the US Economy" (*Wired*, 5 November 2025) https://www.wired.com/story/data-center-ai-boom-us-economy-jobs/ accessed 31 December 2025.

137: "DeepSeek Is Using Banned Nvidia Chips in Race to Build Next Model" (*The Information*, 10 December 2025) https://www.theinformation.com/articles/deepseek-using-banned-nvidia-chips-race-build-next-model accessed 31 December 2025.