Chapter 7

Full Stack AI Governance

Michael Veale and Robert Gorwa

Chapter 7 of *Malice, Models and Middlemen: The Full Stack Fight Against AI Misuse*, pages 176–189

Abstract

In the conclusion, we outline the main lessons and cross-cutting findings from the previous stack chapters. We look at how technoregulation has been distributed across the stack, which has been omitted to date from the regulatory literature on code- or design-based regulation. Our framework of "hooks" and "loops" — the governance residue left by one layer for another layer to connect to and make use of — helps us navigate this and provides a framework for future scholars. We look at the importance of secure enclaves and hardware to governance: physical cryptographic tools which recur through our book. We also consider how friction is crucially important for limiting misuse. Focusing too deeply on one area of the stack typically leads to serious imbalances of power and outcomes, creating disproportionate impacts on people's freedoms, and diminishing returns in service of policy aims. Full-stack AI governance, where many layers are animated at once and interconnect in sensible and considered ways, is much more balanced AI governance. It is not a silver bullet to stopping abuse, but it is a lot shinier than the alternatives.

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 7

https://homepages.ucl.ac.uk/~ucqnmve/malicemodelsmiddlemen/chapter/7/

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.

7 FULL STACK AI GOVERNANCE

FROM SAFETY TO IMPACT AND BACK AGAIN

In February 2026, hundreds of AI researchers trudged through the heat to the Bharat Mandapam convention center, a brightly colored metal and glass swirl that had recently been built, despite much public uproar, on a spot formerly occupied by the architecturally iconic Delhi Hall of Nations building.[1] The AI 'Impact' Summit was in India, the latest in a series of now-annual meetings that tellingly had morphed in focus from their genesis as the UK's AI 'Safety' Summit in 2023.

The researchers were joined by many global policymakers. Prominent industry CEOs like OpenAI's Sam Altman, Google's Sundar Pichai and Anthropic's Dario Amodei rubbed shoulders over dinner with Narendra Modi as well as the heads of state of France, Brazil, Spain and many more.[2] More than 250,000 visitors were reported to have walked through the accompanying public expo hall, checking out booths staffed by representatives of hundreds of startups (both Indian and international) hawking their wares.[3]

The economic import of these firms, and of generative AI tools in general, had massively changed in less than five years. The AI industry was absolutely booming, its rapid rise fueled, at least in part, by massive, circular financial flows between firms, which were eagerly investing in each other to obtain chips and cloud capacity.[4] The financial press was now regularly pontificating not just on the question of whether this constituted a bubble, but on what kind of bubble it was (more like the 2008 global financial crisis, the 1990s dot-com boom or the railroad bubble of the mid-19th century).[5] The chipmaker Nvidia, whose products were in massive global demand and at the center of high-level international negotiations as to which countries should be permitted to purchase them, became the world's first company to achieve a $5 trillion valuation.[6] The market cap of the six most valuable American companies — all

1: Jaideep Deo Bhanj, "'We Are Being Ruled by People Who Are Philistine': Architect Speaks out against Demolition of Hall of Nations" *The Hindu* (20 February 2017) https://www.thehindu.com/news/cities/Delhi/we-are-being-ruled-by-people-who-are-philistine/article17333069.ece accessed 21 February 2026.

2: PTI, "AI Impact Summit: PM Modi Welcomes World Leaders; Holds Bilateral Meetings with 7 Heads of State and 2 CEOs" (*The Financial Express*, 18 February 2026) https://www.financialexpress.com/india-news/ai-impact-summit-pm-modi-welcomes-world-leaders-holds-bilateral-meetings-with-7-heads-of-state-and-2-ceos/4147737/ accessed 21 February 2026.

3: HT News Desk, "AI Impact Summit 2026: All about Global Event in Delhi" (*Hindustan Times*, 16 February 2026) https://www.hindustantimes.com/india-news/india-ai-impact-summit-2026-features-how-to-attend-and-more-all-about-the-event-101771200409940.html accessed 21 February 2026.

4: Cedric Sam and others, "A Guide to the Circular Deals Underpinning the AI Boom" *Bloomberg* (22 January 2026) https://www.bloomberg.com/graphics/2026-ai-circular-deals/ accessed 21 February 2026; Rafe Rosner-Uddin and others, "The Multiplying Risks of Financing Data Centres" (*Financial Times*, 26 August 2026) https://www.ft.com/content/9cb63980-cd43-419d-a7e5-42039b9fd01e?syn-25a6b1a6=1 accessed 26 August 2026.

5: Katie Martin, "Year in a Word: AI Bubble" *Financial Times* (23 December 2025) https://www.ft.com/content/3e411742-b6bb-40c2-9432-4ee6f562af8f accessed 21 February 2026; John Plender, "How the Bubble Bursts" *Financial Times* (3 January 2026) https://www.ft.com/content/7987310a-5c90-4976-b730-3559502006e2 accessed 21 February 2026.

6: Tim Bradshaw, "US Allows Microsoft to Ship Nvidia AI Chips to Use in UAE for First Time" *Financial Times* (3 November 2025) https://www.ft.com/content/03b30ba3-d0c6-4f63-92f8-077fcd8dc472 accessed 21 February 2026; Niket Nishant

of which were tech firms: Nvidia, Apple, Alphabet, Microsoft, Amazon and Meta — was, at the close of 2025, equivalent to about two-thirds of U.S. GDP.[7]

The India AI Summit was also occurring at a wholly different juncture in the AI policy conversation. If the decade beginning in 2010 came to be characterized by a multi-jurisdictional transnational regulatory pushback against the unfettered provision of services being operated by large technology firms around the world, the balance now seemed to have decisively shifted toward letting innovation proceed unencumbered.[8] Chinese leaders were explicitly framing AI development as a central aspect of their future growth strategy, taking intensive steps — such as curbs on exports of rare-earth minerals — to fight back against U.S. chip sanctions.[9] The European Commission began a process of rolling back and "simplifying" its flagship AI Act, negotiated between 2021 and 2024 in the first Ursula von der Leyen Commission.[10] In the United States, a coterie of 'digital oligarchs' was allying itself with the second Trump White House, hoping that the U.S. could once again spearhead their interests globally (in a manner akin to the early days of the 'U.S. Internet Freedom Agenda').[11] Upstart 'non-profit' research organizations had evolved into global mega-businesses spending millions on transnational lobbying, wining and dining policymakers and civil servants around the world.[12]

From an overall governance standpoint, it would be difficult to peer out at the world in early 2026 and see the end result as anything less than a horrible muddle of strategic misdirection, futile grandstanding and inaction. Actors around the world claimed to be engaged in building a governance regime around 'AI,' but what exactly the 'it' was that was being governed, how, by whom and why, was all much less clear.

Some things were clear, however. Thousands of different types of generative models with different capabilities (text-, image-, video- and multimodal) were now easily available to consumers. The most powerful of these systems, the 'frontier' LLMs being developed by

and Rashika Singh, "Nvidia Hits $5 Trillion Valuation as AI Boom Powers Meteoric Rise" *Reuters* (29 October 2025) https://www.reuters.com/business/nvidia-poised-record-5-trillion-market-valuation-2025-10-29/ accessed 21 February 2026.

7: Authors' calculation. Combined market capitalization of Nvidia, Apple, Alphabet, Microsoft, Amazon and Meta on 31 December 2025 from CompaniesMarketCap.com CompaniesMarketCapcom, "Largest Companies by Market Capitalization" (2025) https://companiesmarketcap.com/ accessed 31 December 2025; annualized Q4 2025 nominal GDP from Bureau of Economic Analysis, United States of America, "GDP (Advance Estimate), 4th Quarter and Year 2025" (20 February 2026) https://www.bea.gov/news/2026/gdp-advance-estimate-4th-quarter-and-year-2025 accessed 10 March 2026.

8: Henry Farrell and Abraham L Newman, *Of Privacy and Power: The Transatlantic Struggle Over Freedom and Security* (Princeton University Press 2019); Anu Bradford, *Digital Empires: The Global Battle to Regulate Technology* (Oxford University Press 2023); Robert Gorwa, *The Politics of Platform Regulation: How Governments Shape Online Content Moderation* (Oxford University Press 2024); Fernando Filgueiras, Ricardo Fabrino Mendonça, and Virgílio Almeida, "AI Regulation Versus AI Innovation: A Fake Dichotomy" (*Tech Policy Press*, 18 April 2025) https://techpolicy.press/ai-regulation-versus-ai-innovation-a-fake-dichotomy accessed 21 February 2026.

9: Xinhua News Agency, "Translation of Remarks Made at the 20th Collective Study Session of the CCP Central Committee Politburo" (*Center for Security and Emerging Technology*, 26 April 2025) https://cset.georgetown.edu/publication/xi-politburo-collective-study-ai-2025/ accessed 23 February 2026; Reuters, "China Expands Rare Earths Restrictions, Targets Defense and Chips Users" (*Reuters*, 10 October 2025) https://www.reuters.com/world/china/china-tightens-rare-earth-export-controls-2025-10-09/ accessed 23 February 2026.

10: Daniel Leufer, "Digital Rights Are on the Chopping Block in the European Commission's Omnibus" (*Tech Policy Press*, 19 November 2025) https://techpolicy.press/digital-rights-are-on-the-chopping-block-in-the-european-commissions-omnibus accessed 23 February 2026.

11: Julie E Cohen, "Oligarchy, State, and Cryptopia" (2025) 94(2) Fordham Law Review 563; Shawn M Powers and Michael Jablonski, *The Real Cyber War: The Political Economy of Internet Freedom* (University of Illinois Press 2015).

12: Cam Wilson, "Revealed: How OpenAI Lobbied to Change Australia's Laws — and Got Some of What It Wanted" (*Crikey*, 19 February 2026) https://www.crikey.com.au/2026/02/19/openai-lobbied-change-australias-laws/ accessed 23 February 2026; Cecilia Kang, "How Sam Altman Stormed Washington to Set the AI Agenda" *The New York Times* (7 June 2023) https://www.nytimes.com/2023/06/07/technology/sam-altman-ai-regulations.html accessed 23 February 2026.

major American labs and made accessible in a closed-box manner to paying customers, had capabilities that were actively being used for a wide range of illegal and/or socially deleterious activities. The leading players were seemingly unable to prevent even these relatively low-stakes forms of model manipulation, although they were making — depending on the firm — some degree of investment in trust and safety infrastructures intended to make this more difficult.

But beyond this, there was now a huge long tail of downloadable models that had similar, if not quite as advanced, capabilities. Models being created by small labs, research organizations and even non-commercial entities were easily available, usually for free, to anyone with sufficient know-how and a powerful enough computer (or a credit card to pay for cloud credits). While general-purpose dual-use open models were proving to be quite helpful additions to various bad actors' misuse toolkits, these models were also being widely customized and refined for tasks like deepfake, CSAM or fraud generation. Many of these specialized, overtly malicious models were available on the same app stores and model hosting services.

The activity of the largest AI research labs seemed to be proceeding entirely untrammeled: high-level policy conversations were generally concerned about helping these labs 'win the AI race' or achieve 'supremacy' rather than engaging with AI's short- and medium-term risks. Firms distanced themselves from risks not associated with 'frontier capabilities.' The open model status quo showed no signs of shifting either. International meetings like the one happening in Delhi were instead clearly designed to "create a venue for governments and corporations to announce new ventures and pursue new partnerships," with "one clear economic objective outweighing all others — to seek private sector investment into domestic AI projects."[13]

This isn't to say that absolutely nothing was happening on the regulatory front. Some jurisdictions were beginning to respond to some of the types of immediate model misuse scenarios that we have discussed throughout this book. But these were mostly being pursued in a scattershot, misguided way, targeting the wrong actors and technical infrastructures, and promising to create undesirable incentives that would have unintended consequences. For example, the problems with India's alterations to its IT rules to introduce obligations for cloud model providers and comms platforms around synthetic content, which came just days before its AI Impact Summit, are discussed in Chapters 2 and 6. So what kind of approach would work better?

DISTRIBUTED TECHNOREGULATION

The open-weight model cat is out of the proverbial bag, whether or not private governance or public policy is ready. We are now in a world of downloadable, distributed, expressive AI systems, like it or not. This has involved many different actors, many of which we have discussed in this book. Any types of actually effective governance solutions will also need to be flexible and distributed, to appropriately consider the many intermediaries involved in the provision of these models and the interventions that their position in the stack may allow

13: Amber Sinha, "India's AI Summit Could Prove to Be New Delhi's Lost Opportunity" (*Tech Policy Press*, 18 February 2026) https://techpolicy.press/indias-ai-summit-could-prove-to-be-new-delhis-lost-opportunity accessed 23 February 2026, n.p.

them to take.

There are some who advocate for banning open-weight, downloadable models.[14] Yet this claim needs further specification, and the devil is in the details. Law does not act on things directly; law can only govern how people must act toward one another regarding those things.[15] We might try to set up criminal or civil structures so as to disincentivize open model production and release. This might indeed stop large domestic AI labs releasing them for download. But what about foreign labs? What about criminals using stolen cloud compute or local hardware? What about hacked and stolen model weights transferred through peer-to-peer file sharing? Should these other layers of the stack also be liable? Under what conditions, and would those conditions be effective in facilitating such a 'ban?' Would these provisions, in any case, even be permitted in all jurisdictions, such as under the United States' First Amendment?

It is the premise of our book that, instead of high-concept grandstanding, the actual realities of the harms that AI systems can and will cause will continue to be shaped by a technoregulatory regime that is quietly negotiated in the shadows. This does not mean it will be an entirely private endeavor. Interest groups and policymakers can and will continue to shape global governance of these challenges.[16] Some of these regimes may eventually receive the blessing of law. Others may be challenged by it.

We will come back to the broader question of possible futures in a bit. But first, we would like to draw out the trends and themes, across the stack and the challenges we have studied, that we have observed while researching this book. The issues we have raised in these pages have been complex. This is a dense volume, and we have tried our best to lay out the social, legal and technical challenges as we understand them in an accessible form without forfeiting the complexity that these discussions do need. Yet as readers may have noticed, there have been elements that recurred again and again within that complexity.

DOWN WE GO! OPERATING SYSTEMS AND SECURE ENCLAVES

In some ways, we are living in the world that was anticipated by the Internet law scholar Jonathan Zittrain in *The Future of the Internet — and How to Stop It* almost 20 years ago.[17] That book worried that the efforts to rein in the chaotic, insecure Internet, and the flexible, adaptable personal computers that connected to it, were leading to a future where digital devices were becoming locked down and less customizable. This would come with a high 'generativity' cost, with the danger of losing the core innovative dynamics of highly capable, accessible, easily masterable and teachable computing technologies. Zittrain's book profiled computers

14: See e.g. David Evan Harris, "Not Open and Shut: How to Regulate Unsecured AI" (*Centre for International Governance Innovation*, 6 September 2024) https://www.cigionline.org/articles/not-open-and-shut-how-to-regulate-unsecured-ai/ accessed 26 December 2024.

15: Wesley Newcomb Hohfeld, *Fundamental Legal Conceptions as Applied in Judicial Reasoning and Other Legal Essays* (Walter Wheeler Cook ed, Yale University Press 1919). For the pedants, common law countries have *in rem* jurisdiction, particularly for civil forfeiture cases, where you will find strange case names such as *United States v. 11 1/4 Dozen Packages of Articles Labeled in Part Mrs. Moffat's Shoo-Fly Powders for Drunkenness*, 40 F. Supp. 208 (W.D.N.Y. 1941) or *South Dakota v. Fifteen Impounded Cats*, 2010 SD 50, 785 N.W.2d 272. These are not actually exceptions, but just a customary way of referring to the jurisdiction the court has over the interest of *all persons* in a thing.

16: Michael Veale, Kira Matus, and Robert Gorwa, "AI and Global Governance: Modalities, Rationales, Tensions" (2023) 19(1) Annual Review of Law and Social Science 255 DOI: 10.1146/annurev-lawsocsci-020223-040749 .

17: Jonathan L Zittrain, *The Future of the Internet and How to Stop It* (Yale University Press 2008).

with their generativity stripped away, like the constrained TiVo recording box (much like today's Apple TV) and the Xbox 360 (much like the original Nintendo Entertainment System discussed in Chapter 3). The iPhone and its walled garden App Store, which came into being after the book came out, are the most decisive indication of how the technology industry has dealt with Zittrain's dilemma.

The Future of the Internet helps to explain the rise of today's platformized Web and the capture of much of online life in major tech firms' private fiefdoms. But in other important ways, something not quite foreseen by Zittrain has happened. That book worried that sustained security issues, and perhaps a single calamitous cybersecurity crisis — what Lawrence Lessig, in somewhat dated language, described as 'i9/11' in the book's preface — would lead to a radical remaking of the digital. This dreaded scenario might be caused by a clumsy hobbyist making software that goes terribly wrong, or a state actor deploying malware that spirals badly out of control.

These worries haven't completely faded from view. Perhaps we will yet be ravaged by a devastating agentic system run amok. Perhaps only the printed copies of this book will remain. This is, after all, related to the 'existential risk' framing that at least for a time has been supported by the largest AI firms on the planet. Yet this framing distracts from the harder truth that we laid out in this volume. The insecurity from modern generativity, fueled more than ever by generative AI, is not clumsy and accidental, but malicious and distributed. It is not something that model developers alone are well-placed to do much about, in the same way that the architects of the Internet in Zittrain's book could do little about its eventual misuse.

The contemporary 'AI safety' conversation generally focuses on the model developers delivering the interventions to protect society. Yes, such developers may be morally responsible for at least some of these outcomes. Perhaps they should be asked to pay to remedy the consequences. But when it comes to actually stopping, rather than just adapting to, harm, there is a much wider ecosystem currently in play. Internet policy interventions generally require action by intermediaries, and model developers simply will not always have the visibility or leverage to prevent or even clean up the mess that they have helped create.

In the face of the ensuing chaos, our work supports Zittrain's hypothesis that the device is a major battleground for regulation of generativity. Unsurprisingly, computers matter to computing policy, and operating systems and hardware are devices' foundational components. Yet in the AI policy discourse, at the varying AI summits or at the major AI conferences, operating systems and their developers are rarely considered actors worth talking to. GPU manufacturers are occasionally placed on stage, but as we have discussed, GPUs are extraordinarily badly placed to assess whether the use they are supporting is good or bad. We think that on the basis of the preceding chapters, this is both a huge omission, and a failure of the whole field to fully analyze real, currently happening AI misuse and its trajectories. Operating systems and their associated hardware are presenting technical opportunities for AI governance which are already materializing in many of the parts of the AI stack we have studied in this book. They should be considered as part of these conversations. This does not, however, mean that this form of governance is necessarily the best way forward for society. Operating system–based governance poses enormous challenges for surveillance, control and freedom. Yet even if it is not explicitly discussed, it will not go away. Instead, we risk it happening silently without

political debate, scrutiny or oversight.

In Chapter 3, we showed how operating system control has been pivotal to previous iterations of local software governance. Much of this boils down to software gatekeeping: the boot chain and the chain of trust running from the hardware up to the operating system, which in the case of a platform like iOS, has allowed it to have the last say on what software can run on someone's device. This returned in Chapter 4, where we saw Apple's Private Access Tokens. This was a move promoted by Cloudflare to allow the secure enclave of iOS devices to prove that a device is real and untampered with for the purposes of proving one's humanity — or at least proof that someone has spent a lot of money on an Apple device and had to overcome any security features that might limit scraping on it. As we argued, the difficulties in bot detection mean that this is a likely direction of travel for authentication more broadly. We also saw operating system control at play in Chapter 6, where the C2PA authenticity standard's Level 2 security for signers relies entirely on operating systems' chain of trust to prove that a piece of software digitally signing content it has changed has not been tampered with, and that the keys, again hidden in the secure enclave, cannot simply be extracted and misused by a hacker.

Secure enclaves are controversial. They are often framed as designed to protect users against malicious software that is operating unbeknownst to them on their device, in so-called 'userspace,' such as software that may wish to obtain secret information such as the fingerprints or a facial scan used to unlock their phones. By locking this data in a secure enclave, it should become impossible for other software to touch it. Such protective aims are welcome, as the average user clearly cannot keep track of everything running on their pocket supercomputers, and not everyone wants to be a full-time security engineer in addition to running their own life. Yet it is not just hackers who might lurk in userspace. Users live there too. Cryptographic tools like the secure enclave can also be used to coerce users, preventing them from installing software they want, customizing or tinkering with their devices or the software on them. Such control does not automatically favor the weak or vulnerable, but is associated with moral decisions.[18] During the Covid-19 pandemic, when Apple and Google implemented a contact tracing tool at operating system level, it was designed in such a way, using the secure enclave, that countries had to accept the design schematic the firms laid out, even if they believed — as some did, notably France — that a different approach would be preferable.[19] In choosing to govern AI through operating systems and hardware, any 'digital sovereignty' governments feel they are gaining against the challenge of AI misuse has to be balanced against that which they are ceding to Apple, Google, Microsoft, Nvidia and friends.

Operating systems are also a vantage point for surveillance. Your operating system knows what you are using your device for. If you cannot trust your operating system, it is very hard to say anything about the security of your device. That is why operating systems are such

18: Phillip Rogaway, "The Moral Character of Cryptographic Work" (*Essay accompanying the IACR Distinguished Lecture, AsiaCrypt 2015*, 2015) https://perma.cc/KV9S-B7LJ; making related arguments on the gatekeeper power enclaves facilitate in cloud environments, see Jatinder Singh and others, "Enclaves in the Clouds" (2021) 64(5) Communications of the ACM 42 DOI: 10/gjx2dz.

19: Michael Veale, "Sovereignty, Privacy and Contact Tracing Protocols" in Linnet Taylor and others (eds), *Data Justice and COVID-19: Global Perspectives* (Meatspace Press 2020); Tamar Sharon, "Blind-Sided by Privacy? Digital Contact Tracing, the Apple/Google API and Big Tech's Newfound Role as Global Health Policy Makers" (2021) 23 Ethics and Information Technology 45 DOI: 10.1007/s10676-020-09547-x.

alluring targets for intelligence agencies to develop or purchase spyware to compromise, such as NSO's Pegasus or GCHQ's SMURF suite. Once your operating system is compromised, you have little hope for your personal privacy or security. We saw the potential power of operating system surveillance in relation to AI misuse in Chapters 3 and 6. Operating systems can literally monitor your screen and run classifiers to see what you are doing, either what you are generating, or what you are receiving. Smartphone companies have already integrated such technologies to try to detect scams and deepfakes, with varying levels of success.

To date, legal mandates for operating system surveillance have played out in legislative pressures for smartphone providers to implement technologies for detecting child abuse images and grooming activity. This is the debate on client-side scanning, known by critics as 'chat control.' While there are issues about whether or not this would even work at all, or whether it would be helpful, the bigger political challenge is the encroachment of the state into more and more private and intimate spaces. Technoregulation can bake surveillance and control into a device, but it cannot do so in a way that guarantees the technology will not be misused. As is often said by technology executives, you can't have a back door that only the good guys can walk through.[20]

DIVIDE AND CONQUER! SEGMENTATION STRATEGIES

Another theme running through this book is the segmentation or segregation of the governed entities to create a workable regulatory load. When pushing governance at almost any stack layer to its limits — not always a great or desirable idea for society — we typically find that such moves create other, unintended problems. Often, the only way to make intense governance at a single layer at all tractable is if the actors being governed are separated out such that different modalities of governance can be applied to different groups.

In Chapter 2, we concluded that one potential way out of cloud governance difficulties, given the challenge of mitigating or even surveilling the risks posed by unmanaged compute, would be to segment cloud users into (1) licensed users, who would have full access to cloud services but be governed by organizational controls, and (2) unlicensed users, who would only be given access to the easy-to-monitor tools.

In Chapter 3, we described how, in the context of local governance, given the constraints that operating system control would place on research and development and the difficulties in applying it to the lowest-level compute, segmentation would also be needed. Consumer devices, were they to be locked down and filled with technoregulatory filters and scanners, would have to be distinguished from research-grade systems, which would have to remain open, and controlled through supply chain regulation to ensure that only authorized actors could use them. Such segmentation, we observed, already exists in other areas of law and technology, such as markets for drones, or the regulation of amateur radio. We also saw the

20: Meredith Whittaker, "Online Safety Bill Debate: Could It Lead to 'Unprecedented Paradigm-Shifting Surveillance'?" (3 July 2023) https://www.channel4.com/news/online-safety-bill-debate-could-it-lead-to-unprecedented-paradigm-shifting-surveillance accessed 10 March 2026, (president of Signal); Danny Yadron and Daisuke Wakabayashi, "NSA, Apple Chiefs Decode Encryption Views - WSJ" (*Wall Street Journal*, 20 October 2015) https://www.wsj.com/articles/nsa-apple-chiefs-decode-encryption-views-1445318890 accessed 10 March 2026, (comments by Tim Cook, CEO of Apple).

potential for model notarization, which may allow identified models, or models certified to pass certain tests, to be treated differently from non-identified or certified ones.

In Chapter 4, we observed how, at the network layer, bot detection was failing so badly that network actors such as Cloudflare and Google were moving to segment the problem in multiple ways. One was to consider 'who has a real, untampered-with device' compared to who does not, attempting to use browser tests and hardware attestations. Those that do not were further segmented into good, 'verified' bots that could authenticate using credentials and promise to follow the rules of network intermediaries. Rulebreakers would fall into the other category: bad, and ultimately blocked, bots.

In Chapter 5, we saw how platforms like Hugging Face struggle with takedown requests against models. Laws designed to immunize such platforms often trigger when they are notified of illegal content on their service, but the illegality of models is neither technically easy to check nor to confirm. This leads to a dilemma for model hosts — either over-moderate when a notification is given due to the sheer difficulty of checking its accuracy, or under-moderate and risk eventual liability. Such dynamics can easily be abused by those using notification systems to try to force takedowns with mischief or commercial gain in mind. One practical solution to this could, again, involve segmentation, allowing uploaders that meet some kind of due diligence standard to have the case for keeping their models up considered more carefully by the platform, while users lacking these checks may find their uploaded AI artifacts more easily removed when they are complained about.

Finally, in Chapter 6, we described the different paths forward for 'authentic' content on comms platforms like social networks. When users upload 'authentic' content, such as that recognized by different levels of C2PA-signed metadata, platforms face choices about how to segment out governance. Should content with different levels of authorization face different labeling? Be boosted or suppressed in recommender systems? Should it be more quickly taken down when reported?

These segmentation strategies are never exactly the same across the stack, but in most instances share a common thread: they help mitigate the unbearable externalities of inherently clunky technoregulatory mechanisms. Ruthlessly block all bots, and you block useful automation and the Internet Archive. Infantilize all cloud or on-device computing into plugging together restricted and monitored APIs, and you sacrifice scientific progress and innovation. Judge all content by its authenticity metadata, and you lose the important historical or expressive content that is unauthenticated, as well as content from those whose devices cannot, or have good reason not to, leave behind identifiable trails of who they are. Technoregulation is powerful, but it is not flexible. When it does work, it only deals well with narrow problems. Issues seemingly as narrow as those tackled in this book — we, after all, haven't been talking about other thorny AI policy issues, like bias or discrimination — need to be whittled down before technoregulation can effectively be applied. Wield those governance mechanisms too wildly, too ambitiously or with too much hubris, and they will surely backfire.

SLOW DOWN! FRICTION FOR THE MASSES

We've discussed a huge number of different use cases, and different technologies all generally lumped in under the generative AI moniker. Disrupting all malicious uses of all these different tools, everywhere, all the time, is, given the nature of the globally distributed Internet and the devices that run on it, inherently impossible.

Perhaps that is alright. Some insights into this can be gleaned from the political science scholarship on information controls, and in particular, the controls imposed by the widely acknowledged world leader in online censorship, China. It is a common misconception that the Chinese government tries to permanently 'ban' information and make it inaccessible. China's blocking techniques are always intentionally porous and imperfect. They strike trade-offs through their imperfection. Rather than crude censorship, some scholars have re-conceptualized these techniques as "a *tax* on information, forcing users to pay money or spend more time if they want to access the censored material."[21]

The aim of Chinese information controls is thus to impose friction that will make the average user less likely to obtain certain materials, rather than exerting perfect control. The blocking imposed by Chinese ISPs and platforms can be worked around with VPNs, and users rephrase and paraphrase posts to avoid removal, meaning that people willing to do a bit of searching can generally find what they are looking for. The idea is to deter this kind of activity, making it more annoying and time-consuming, with the underpinning assumption that — except at particularly salient moments of political crisis — the average member of the public will simply not bother.

This logic assumes that the demand of citizens for political information is highly elastic. Small changes in the cost of obtaining it will have a correspondingly high effect on the probability of them actually seeking it out.[22] It's a yet-to-be-examined empirical question to what extent this is true for the kinds of AI harms we have explored in this book. And policymakers and other governance stakeholders in high-income democratic contexts will surely recoil from comparisons between their well-intentioned efforts and the actions of the Chinese government. But friction is an important tool for doing difficult balancing acts in technological policy more generally, for instance when it comes to maintaining both people's right to obscurity online and the public's ability to learn and access information freely.[23]

Friction can most obviously be applied to contexts that could impact the ease of learning about, obtaining and using illicit customized versions of general-purpose AI models.

In Chapter 5, we examined the interaction between app stores hosting 'nudification' apps and other platforms that featured advertisements for those apps. This should be a relatively easy issue to address, given the widely reported trend of young people using AI tools to create non-consensual intimate imagery of their classmates, usually young girls.[24] Emerging on-the-ground research in these contexts has pulled out the importance of usability in this specific

21: Margaret E Roberts, *Censored: Distraction and Diversion Inside Chinas Great Firewall* (Princeton University Press 2018) 2.
22: ibid 31.
23: Woodrow Hartzog, *Privacy's Blueprint: The Battle to Control the Design of New Technologies* (Harvard University Press 2018).
24: Elizabeth Laird, Maddy Dwyer, and Kristin Woelfel, "In Deep Trouble: Surfacing Tech-Powered Sexual Harassment in K-12 Schools" (*Center for Democracy and Technology*, 26 September 2024) https://cdt.org/insights/report-in-deep-trouble-surfacing-tech-powered-sexual-harassment-in-k-12-schools/.

area, noting the inherent barrier posed by the technical skills required to download, prompt and potentially modify downloadable models:[25] while some of the most motivated, troubled youth may indeed go to great lengths to try to create these kinds of images, for others, friction in terms of lacking easy access to the software, technical skills or tutorials may be enough to deter them. After all, this kind of image editing is not new, just easier and more powerful — motivated youths could have gotten similar results with Photoshop for decades. It would be trivial for large search engines and app stores to make it more difficult to find these apps, which remain easily accessible via leading providers,[26] by delisting certain results and by blocking certain search queries. A quick examination in February 2026 from our European IP addresses indicates that Google's autocomplete results have been disabled for 'deepfake porn' but still suggest 'deepfake maker AI clothes changer' and 'deepfake app online no login free.'[27]

Other intermediaries, such as Web hosts and network services, may well be mobilized to take down the technologies that are easiest to use and abuse. We already outlined in Chapters 1 and 5 how model hosts have often removed one-click deployment options aimed at less tech-savvy users before fully taking down a model. The type of signals sharing we looked at in Chapter 6 may well also have a role in identifying what should be a target for governance, and may highlight and orchestrate roles for intermediaries across the stack — cloud APIs being abused; abusive models being hosted; RAGs and agents being connected to these services.

Tutorials, as we looked at in Chapter 6, can be trickier to add friction to, as they are often framed in dual-use terms, and it may be legitimately difficult to justify removing them from comms platforms. Yet the 'last mile' of generic tutorial use is to turn something generic into something for your specific tasks. Enter LLMs again: they promise to write software for us, and to transform current software developers into shepherds for little herds of coding models. Whether that exactly comes to pass or not, it seems likely that people will become more reliant on LLMs for getting computers to do what they want, even if only because they don't know how to code themselves anymore and struggle to learn in a world of obvious shortcuts and lower demand. This places these coding assistants as another location for friction and governance.

AI literacy is wheeled out as a trite recommendation when policymakers, think tanks, lobbyists and academics have nothing left of interest to say in a report. Similarly, our call for what you might frame as a type of AI *illiteracy* (at least for the bad guys) will only take us so far. Fraud, phishing and scams have always been part of online life, and it doesn't seem that the demand for effective AI scam tools will be highly elastic. Those determined to commit fraud effectively have strong economic incentives to do so, and there will always be motivated scam gangs and highly skilled individuals tapping into the dark arts of making money at a distance. Yet even in these cases, friction added elsewhere may make a difference. Authenticity

25: Riana Pfefferkorn, "Addressing AI-Generated Child Sexual Abuse Material: Opportunities for Educational Policy" (*Stanford Human-Centred Artificial Intelligence*, 21 July 2025) https://hai.stanford.edu/policy/addressing-ai-generated-child-sexual-abuse-material-opportunities-for-educational-policy.

26: Alexios Mantzarlis and Santiago Lakatos, "AI Nudifiers Continue to Reach Millions and Make Millions" (*Indicator*, 14 July 2025) https://indicator.media/p/ai-nudifiers-continue-to-reach-millions-and-make-millions; Valerie A Lapointe and others, "The Governance of AI-generated Pornography Platforms: A Content Analysis" [2026] New Media & Society 14614448261421873 DOI: 10.1177/14614448261421873 .

27: For more on Google search moderation, see for example Richard Rogers, "Algorithmic Probing: Prompting Offensive Google Results and Their Moderation" (2023) 10(1) Big Data & Society 20539517231176228 DOI: 10.1177/20539517231176228 .

standards like C2PA or similar built into video calls, and discussed in Chapter 6, may not be insurmountable for scammers, but may be proportionate for video calls and voice notes, and mean that a scammer has to at least explain away a warning label every time, or move people suspiciously to a platform that does not require such a connection. Even if this were not obligatory or even default, it may give family members the option to enable enhanced security on more vulnerable relatives' devices to try to avoid them falling for such scams. The signal sharing we discussed may force criminals deeper underground into peer-to-peer or darkweb platforms to avoid discovery and disruption. This may feel like a loss, but these places are by definition harder to use, harder to find and can require a lot of operational security knowledge to operate within without slipping up.

The network layer in Chapter 4 has some of the most effective friction-based governance. Scraping around Cloudflare is not impossible, but it is expensive, slow and requires a mountain of consumer hardware that could ostensibly be operated by a real person, rather than just a single machine, or a hijacked cloud server. This has not removed bots from the equation, but the alternative, where they can be spun up ad infinitum, is likely much worse than the situation we have today.

Friction will be a key tool in the future AI policy toolkit. Friction may not remove the threat entirely — it is a force that slows rather than one that stops. But when evaluating the efficacy of interventions, it is important not to miss the fact that without it, these problems could become much, much worse.

HOOK THAT LOOP! SEEING AND ACTING ACROSS THE STACK

Misuse happens across the AI stack, but stack actors are often working in isolation from each other. When they work in isolation, they end up at the logical endpoints of each of our chapters — failed technosolutionism, hubris, despair. There is only so far each layer of the stack could, and likely should, go. Instead, greater emphasis needs to be placed on working together.

The first step is seeing across the stack. For some harms, this is definitionally crucial. In Chapter 6 we observed that while some harms (like child abuse material) are detectable anywhere, other harms are mainly perceptible in context, and there comms platforms perform a uniquely important flagging role. In general, signals sharing platforms like Lantern are going to be more important than the classic hash-sharing platforms like GIFCT for terrorist content or NCMEC for child abuse, given that AI models extrude never-before-seen content which hash-based techniques struggle with.[28]

There are various governance loops that some stack actors could also create that others could hook into actively, during the course of investigation. Where one stack actor finds, for example, an interactive system (like a scammer chatbot), they would not usually be able, on their own, to figure out the source of the AI or computation. It is highly likely such content is not posted directly from the IP address of the computational resource (like the AI model or generic compute provider), but via a series of intermediate actors, servers and proxies. If

28: Some other relevant initiatives in this regard in non-AI-specific areas include the Amazon Anti-Counterfeiting Exchange (ACX) and the Messaging, Malware and Mobile Anti-Abuse Working Group (M3AAWG).

cloud providers left a governance loop that could be used, such as a way to request a search of their logs for certain prompts, even if just logs they retained within a very small time window (a so-called 'quick freeze' approach to data retention), they could unearth malicious actors.[29] Similar quick freeze approaches applied to non-AI services may be useful where models query those services via RAGs or agentic tools, which may reveal the IP address of the server, or of the local machine, doing the computation.

Model providers could also train their models, or otherwise tweak their broader software pipelines, to have secret prompts or triggers which, when used, could act as a governance loop. This would allow a downstream actor with query access to a malicious pipeline (such as if they posed as a victim) to use that hook prompt to identify the underlying model and computational resource. Open-source model providers could train parallel loops, such as keywords or prompts which reveal information the model processes but does not usually include in its output, such as the system prompt, which could in turn reveal useful information about the abuser's setup: URLs, API keys, details about specific RAGs and tools used.

We have also seen that certain stack actors can attempt to leave loops, such as watermarks and metadata, which heighten observability later in the misuse pipeline. Such markings may be simply designed to indicate that something is artificial or, as per requirements in Indian and Chinese law, to actually identify the specific computational resources used. Sophisticated actors may be able to remove these loops, but the existence of such watermarks may not be self-evident, and scrubbing methods may not always work, may be forgotten or omitted (as we saw in Chapter 2) or may themselves leave a suspicious trace which itself could act as a governance loop. Notarized models, as we saw in Chapters 3 and 5, also leave a loop for downstream actors to hook onto.

Even simple information sharing can be a governance loop. Dual-use apps that are sold on an app store as innocuous, but promoted through the advertising ecosystem for malicious purposes, do not have clear reporting chains linking these two. When someone reports a malicious ad, that ad platform should be able to (or be forced to by law) hook into a governance loop in the relevant, hyperlinked, app store record to transmit the relevant information. Individuals and NGOs have been given more rights to report to platforms and trigger takedowns, such as through the 'trusted flagger' mechanism of the Digital Services Act, but perhaps they should be given rights to report directly into signal-sharing repositories, too.[30]

Such coordination will look different for every issue. For any specific AI use case causing a social problem, this problem will have to be analyzed in detail; the technical functionalities and the social realities properly understood; and the intermediaries that are relevant identified and brought to the table. After that information has been identified, and with those actors and other representatives of societal interests in discussion, then creative options and interconnections may start to emerge.

There are huge remaining issues here that will need to be resolved. Who pays? Who coordinates? Who is accountable? What jurisdiction does this even happen in? We leave

29: Case C-140/20 *GD* ECLI:EU:C:2022:258, para 26; see, for a similar proposal in the case of AI-powered criminal essay mills, Noëlle Gaumann and Michael Veale, "AI Providers as Criminal Essay Mills? Large Language Models Meet Contract Cheating Law" (2024) 33(3) Information & Communications Technology Law 276 DOI: 10.1080/13600834.2024.2352692.

30: Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC [2022] OJ L265/1 (DSA) art 22.

detailed questions of such institutional design for future work and discussion. In any case, it will involve a mix of puzzling, powering and participation — all with the aim of working toward agreement on what is known about the problem, and the norms and values at stake from intervening in it in different ways.[31] It will not be easy work. But it will be crucial work, with concrete and immediate near-term impacts.

FULL-STACK AI GOVERNANCE IS BALANCED AI GOVERNANCE

Some readers may think that this book sketches out a recipe for breaking the open Internet. We disagree. The biggest threat to the open Internet from AI is rather that various actors responding to inevitable AI misuse are forced — whether by economic or legal might — to internalize a disproportionate response within their level of the stack alone. As we have shown throughout this book, that is a recipe for peril. The diminishing returns to the potential to disrupt misuse by working at one layer of the stack push actors toward a totalizing form of surveillance and control. Technoregulation is useful, but when pushed, becomes technosolutionism, which both fails at its aims and leaves scorched earth in its wake.[32]

Thinking about AI policy in terms of layers, loops and hooks allows us to start thinking about a tangible toolkit for intervention and coordination. It permits a clearer delineation of responsibilities, and clearer accountability for platforms that choose not to provide legitimate, useful, proportionate loops that can be used for coordinated abuse prevention. It fundamentally puts more options on the table for policymakers, and is our best hope of avoiding an ensuing doom spiral of demands to wreck encryption, surveil everything and crush the last vestiges of computational generativity.

Full-stack AI governance also distributes the costs and decisional power of regulating misuse. A significant criticism of modern platform regulation is that it effectively makes platforms 'too big to fail.' By governing through them so much, we are losing any ability to govern without them.[33] It is true that we do require intermediaries to govern networked technologies: we have illustrated that by placing AI regulation in historical context throughout the book. Yet here, AI is actually different. While you might be able to regulate, at least to some extent, many speech-related issues through piling enormous responsibilities onto a small number of social media platforms at the expense of creating monopolies and the kind of power dynamics that help Big Tech stay big, it isn't possible to pile responsibilities onto a single type of intermediary to effectively stop AI misuse. In the AI stack, the harms are distributed, as are the points of visibility and leverage. The current dominant strategy of modern platform and AI regulation, which involves a broadside against sectoral giants, simply will not work.

Learning to distribute roles and responsibilities, to promote coordination and cost-sharing, is a safer bet. Even though more actors are involved, and more power can be wielded by some players in some cases than others, it could be, if done right, much better for Internet freedom and civil liberties online. Without explicit mechanisms and incentives to shape the ensuing dynamics of distribution and coordination, and without the creativity from working

31: Robert Hoppe, *The Governance of Problems: Puzzling, Powering and Participation* (The Policy Press 2010).
32: Ben Green, *The Smart Enough City* (MIT Press 2019).
33: Cory Doctorow, *The Internet Con: How to Seize the Means of Computation* (Verso 2023).

together and seeing the full array of possible options, individual stack actors will be forced to governance extremes. Some, like operating system providers, might leap there gladly to seize power and attempt to make themselves even more indispensable. Others, like leaner hosting players, will do everything they can to avoid taking on more responsibilities, worried that their action will simply lead to a growing mountain of regulatory demands.

In the face of a multitude of genuine, serious risks from technologies that actually exist today, the way to tackle these risks is to foster wider options for more balanced forms of governance. Defending online freedoms and diffusing the power of private behemoths will require creative, distributed alternatives that can be, on balance, as good as or better than the seemingly simpler, totalizing, single-layer forms of technoregulation. Full-stack AI governance is that balanced governance.