Chapter 1

Introduction

Michael Veale and Robert Gorwa

Chapter 1 of *Malice, Models and Middlemen: The Full Stack Fight Against AI Misuse*, pages 1–18

Abstract

This chapter sets up the book's scope and approach. We argue that in order to understand policies and enforcement in the AI stack, we need to pay much more attention to a much wider array of intermediaries. Rather than focusing on the big AI labs training models, or on evaluations of frontier systems, we examine the past, present and emerging practices of many of the intermediaries which malicious users of AI interact with and through today: cloud providers, operating systems and hardware, network services, model hosting platforms, communications services and more. Code-based "technoregulation," private governance and law in this area are fast growing and entwining. Trying to animate any single layer of the AI stack to fully achieve a complex policy aim against motivated, malicious users of these computational systems usually brings far too many trade-offs to rights and freedoms, if such an approach even works at all. More promising is what we call full-stack AI governance, where interventions are spread across many actors, coordinating to leave technological "loops" — governance residue like watermarks, logs, hashes and blocking capabilities — that other stack actors can "hook" into. We outline the book's structure and approach — looking back before the generative AI rush to historical analogies that bring significant lessons for today, while deeply probing practices and research we see in the current moment for lessons for the future.

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 1

https://homepages.ucl.ac.uk/~ucqnmve/malicemodelsmiddlemen/chapter/1/

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.

1 INTRODUCTION

Close your eyes for a moment and try to visualize a baby daikon radish in a tutu. Or an avocado-shaped armchair. Or a snail, sitting on a pocket-sized harmonica.

These are the kinds of use cases that researchers at the San Francisco-based company OpenAI presented in 2021 when announcing their release of a new image creation system, known before long as DALL·E.[1] As the firm explained in a short press release, "just as a large transformer model trained on language can generate coherent text, the same exact model trained on pixel sequences can generate coherent image completions and samples."[2] The results were cutesy, if lo-fi, images. At 64×64 pixels, they were lower in resolution than those taken by a 1998 Game Boy Camera.

More than five years later, generative AI tools that output text, images or video are ubiquitous. We are now living in a world where one can, without writing a line of code or even paying a cent, easily query an AI system on a phone or laptop to create competent animations, graphics and emojis — harmless illustrations suitable to be displayed in a children's book or pinned to a community center corkboard.

But we are also now living in a world where these systems can be used to do a lot of damage. The very same genAI tools creating innocuous graphics can also create non-consensual sexual content, generate synthetic child sexual abuse material or facilitate new forms of personalized fraud. Text, including text designed to deceive and entrap, has also become similarly cheap.

Take fraud, for example. Scammers are rapidly adopting AI tools and integrating them into their workflows. In some countries, it is common to see faked videos of trusted pharmacologists, luring users into handing over their banking details by promising exclusive new medications.[3] Organized crime groups are making use of voice cloning technology to successfully scam executives at large firms.[4] Online romance scams use open-source face-swapping and similar tools in scam call center compounds.[5] Language models are used to better profile, target and exploit victims.[6] Billions of dollars have been flowing through illicit marketplaces, such as Cambodian conglomerate Huione Guarantee, where AI tools for scammers change

1: OpenAI, "Image GPT" (17 June 2020) https://openai.com/index/image-gpt/ accessed 27 July 2024.

2: ibid.

3: Linda Nordling, "Scientists Are Falling Victim to Deepfake AI Video Scams — Here's How to Fight Back" [2024] Nature DOI: 10.1038/d41586-024-02521-3.

4: Daniel Thomas, "WPP Boss Targeted by Deepfake Scammers Using Voice Clone" (*Financial Times*, 10 May 2024) https://www.ft.com/content/308c42af-2bf8-47e4-a360-517d5391b0b0 accessed 13 August 2024.

5: Will Jackson, "From Face-Swapping to Chatbots — Here's How South-East Asia's Scammers Are Using AI Tools" (*ABC News*, 15 May 2024) https://www.abc.net.au/news/2024-05-16/pig-butchering-scams-artificial-intelligence-ai-face-swapping-/103804830 accessed 13 August 2024; Matt Burgess, "The Real-Time Deepfake Romance Scams Have Arrived" (*Wired*, 18 April 2024); United Nations Office on Drugs and Crime, "Transnational Organized Crime and the Convergence of Cyber-Enabled Fraud, Underground Banking and Technological Innovation in Southeast Asia: A Shifting Threat Landscape" (*UNODC*, October 2024) https://www.unodc.org/roseap/uploads/documents/Publications/2024/TOC_Convergence_Report_2024.pdf; 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; 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.

6: United Nations Office on Drugs and Crime (n 5) 10.

hands.[7]

These are just a few examples, from one issue area. There are many other ways that people around the world are adapting the latest and greatest AI tools to extort, cheat, defraud, abuse and otherwise harm other people.

This book is not about how we got here. It is not about the political, economic, or even technological drivers of these developments. Rather, it's about what happens next. Now that these technologies exist, can any of these problems be mitigated? If so, by whom? How? And at what cost?

WHO CONTROLS THE STACK?

Consider a future where lawmakers in one or more countries come together and agree that there are some things that people should not be able to do with AI tools. Perhaps they should not be able to create realistic-looking intimate images of ex-partners or of Internet influencers. They should not be able to clone the voices of celebrities or ordinary individuals for scamming. They should not be able to use sexually explicit videos of children to create new scenarios depicting those same victims.

But how would this kind of policy be achieved? It isn't entirely clear where one would start. Consider the number of entities that might be involved in a typical modern image generation model, like the 'Stable Diffusion' series, an openly available alternative to DALL·E. The system might be based on a large foundational model (1); which was distributed or hosted online on a platform (2); trained using data scraped by another organization (3); lifted from user content on various Web services and apps (4); before being labeled and cleaned by contracted workers around the world (5). This foundation model might be specified and fine-tuned using a mix of synthetic and specialized data sources (6); using cloud computing infrastructure to do the computation (7); re-hosted on another platform (8); and re-downloaded and used by an individual (9) on their own operating system or device (10). Running the model might be facilitated by specialist machine learning chips (11) and openly available software toolkits (12). There could even be a facilitating entity involved in stringing all the above together and selling it as a service, perhaps as a Web or mobile app (13).

The AI policy spotlight has to date been sharply focused on big developers such as OpenAI or Anthropic. But no matter the pressure placed on them or the responsibilities they are handed, the simple fact is that these firms are not always well-placed to stop the wide range of malicious activities that the technologies they produced are now facilitating. Some AI models that such labs release run 'locally' on individuals' devices, making them much more difficult to monitor and filter. Motivated actors 'uncensor' models and disseminate the means to do so, evading even the best efforts of AI firms to build 'guardrails' and filters around their systems. Tasks undertaken by malicious AI users might look very similar to *bona fide* tasks

7: Andy Greenberg and Lily Hay Newman, "The $11 Billion Marketplace Enabling the Crypto Scam Economy" (*Wired*, 10 July 2024) https://www.wired.com/story/yahoo-boys-real-time-deepfake-scams/; Elliptic Research, "Huione Guarantee: The Multi-Billion Dollar Marketplace Used by Online Scammers" (*Elliptic Research*, 9 July 2024) https://www.elliptic.co/blog/cyber-scam-marketplace; Sue-Lin Wong, "Scam Inc.: 6. Buckets of Billions" (*The Economist (podcast)*, 6 February 2025) https://www.economist.com/podcasts/2025/02/06/6-buckets-of-billions accessed 20 February 2025; United Nations Office on Drugs and Crime (n 5) 9.

that should not be blocked. Sometimes, when taken too far, the legitimate surveillance of AI users' behavior may veer into illegitimate mass surveillance of the type that chills, controls and becomes ripe for abuse.

This doesn't mean that today's AI giants are not the actors to be held morally or financially culpable. We'll leave that to ethicists and litigators. Our point is simpler: such firms *literally cannot* be relied upon to disrupt some of the worst forms of AI misuse. Put another way, legitimate societal worries about malicious deepfakes, spam or disinformation output by LLMs and other generative models cannot be remedied by only foisting responsibilities onto AI's biggest players.

Instead, we need to start thinking about *intermediaries* — the numerous entities involved in the broader ecosystem of technologies that come together to make these complex AI systems functional today. These actors play various roles in what we call the AI *stack*. Sometimes they have enough power to spot and block misuse. Sometimes they can flag it for others to tackle. Occasionally they might be able to leave some governance residue behind, such as watermarks attached to synthetic content, or logs of attempted generation; in certain cases, they might be well-placed to hook onto that residue and use it to mobilize further action.

These measures inevitably vary in effectiveness. They also vary in the magnitude of their side-effects, which can threaten everything from technological openness to individual privacy. We do not write to recommend any specific measure. Many of them have proven themselves bad ideas in the past. In this book, our aim is to help those who inevitably will try to stop abhorrent AI-facilitated activities approach the messy reality of doing so — with all its trade-offs and all its consequences — with their eyes open.

GOVERNING GENERATIVE GENERATIVITY

This is a book about governance in the world of distributed, expressive, generative AI. The questions at the heart of our analysis are complex but not entirely new. How are companies, governments and other actors in the AI policy ecosystem responding to both current and prospective misuse of generative AI? How well do these interventions function as policies? How realistic are they technically? What are the shortcomings and knock-on effects of these developments? There are many examples of ostensibly 'impossible to regulate' or ungovernable networked technologies eventually reaching some kind of policy equilibrium following heavy pressure from governments, civil society, powerful corporations and other stakeholders. The current wave of AI might seem new, but the issues it is raising have striking parallels in the past.

In the mid-2000s, the legal scholar Jonathan Zittrain famously outlined the concept of *generativity*, arguing that it was at the heart of what made computing technologies so powerful and useful.[8] In his terms, a technology is generative if it brings high capacity for leverage over a range of different tasks, is accessible and easy to master. Such generativity comes with a trade-off. The same productive, innovative potential that allows someone to produce useful software on a cheap personal computer in their house simultaneously creates security risks

8: Jonathan L Zittrain, "The Generative Internet" (2006) 119 Harvard Law Review 1974; Jonathan L Zittrain, *The Future of the Internet and How to Stop It* (Yale University Press 2008).

and manifold policy problems, such as malware, distributed-denial-of-service (DDoS) attacks and spam. The problem of malware provides perhaps the most famous illustration of these tensions: the proverbial hacker-in-the-bedroom can write code that can compromise millions or even billions of devices across the world.

Generative AI systems are typically called such because they generate new media, usually after being queried with text prompts entered via a user interface. But they're also generative in this Zittrainian sense: their innovative potential simultaneously facilitates a huge range of potentially damaging applications. Like computing or social media, which have been understood by a socio-technical research tradition as ambivalent dual-use technologies,[9] generative models have technical capabilities permitting not just benign or even beneficial everyday uses, but also uses far more damaging, harmful, surveillant or even violent. As years of technology policy scholarship have taught us, curbing the harmful aspects of these kinds of technologies is inherently difficult. All governance interventions have consequences, and can substantially limit the very generativity that makes these kinds of technologies both transformative and desirable.

Looking back at many years of scholarship on the history of the Internet, software and platform regulation in particular offers a few lessons that we have drawn upon to inform our analysis, and that have shaped the scope and focus of our book.

First, when it comes to disrupting abuse on the Internet, intermediaries have always played an outsized role. As we have already begun exploring above, there are many reasons to expect that this will continue to be the case when it comes to the malicious use of AI models.

Second, the private decisions of these actors when reacting to policy problems — sometimes called self-regulation, regulating with 'code,' or informal intervention — have outsized impacts. What we refer to throughout the book as 'technoregulation,' a mix of private policymaking, technical rule-making and system design, can have bigger impacts than state laws do. Private governance is not just a way of implementing design-based regulation. In many cases, the fact that a firm shows it can control something, even for its own ends, emboldens states to lay more responsibilities at its door. Too much discussion of AI policy focuses on state-level declarations or the growing number of AI-related legal frameworks, and too little focuses on the actual technoregulatory realities on the ground that will shape how a demand from a legislature or executive body can be translated — or not — into reality.

Third, when thinking about the risks to be mitigated, we shouldn't center the fantastical at the expense of the inevitable. Early Internet regulation scholars worried deeply about online catastrophes, particularly devastating cyberattacks, that, once released into the wild either accidentally or intentionally, would permanently cripple the Internet's infrastructure.[10] We don't dismiss unlikely long-term risks entirely. But we should be clearly attuned to the everyday, already-here policy problems that tend to decisively shape how a technology is governed over time. The idea that each breath spent discussing near-term, realized harms is a breath that detracts from the tackling of speculative, prospective future harms is ludicrous.

9: Whitney Phillips and Ryan M Milner, *The Ambivalent Internet: Mischief, Oddity, and Antagonism Online* (Polity 2018); Jonathan B Tucker, *Innovation, Dual Use, and Security: Managing the Risks of Emerging Biological and Chemical Technologies* (MIT Press 2012); Peter Henderson and others, "Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models" in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '23: AAAI/ACM Conference on AI, Ethics, and Society, ACM 8 August 2023) DOI: 10.1145/3600211.3604690 .

10: See for instance Lawrence Lessig's preface to Zittrain, *The Future of the Internet and How to Stop It* (n 8).

There is only so much that can be done about the future, and there is plenty of work to do about the present.

These three arguments relate to three decisions that we have made about the scope of this work, all of which we believe distinguish this book from the rapidly growing number of other publications on AI policy and governance. We'll flesh them out a little more substantively in a moment, but first, let us quickly discuss how we have organized the rest of the book and offer some ideas as to how you might make use of it.

HOW TO READ THIS BOOK

We've written this book for a mixed audience of practitioners, policymakers and academics. We have sought to make our work accessible, and done our best to avoid digressions into conceptual and theoretical matters that may be of interest to scholars but are less clearly relevant for others. That said, much of the material presented in the ensuing chapters is necessarily quite technical. Some is quite dense. AI systems are complex software; to understand how an actor may be able to screen and filter its outputs, or watermark content, or develop an effective new logging regime, requires peeling back the curtain on all that software engineering and looking carefully at how the system, as well as the desired intervention, actually works.

The inherent complexity of the subject matter means that the book will be most effective if the chapters are read in sequence, and in concert. We have focused each ensuing chapter on one part of the AI ecosystem, discussing a different stack 'layer' and set of actors. As these layers build upon each other, and some practices recur in slightly modified forms across the book, readers will be able to understand the more complex concepts if they stick with the text. For instance, Chapter 3 and Chapter 5 both describe different sides of control of mobile phone apps, and each reinforces the other. However, while each chapter builds upon the last, we've also done our best to craft chapters so that they can be useful as standalone reading, and can be looked at in isolation if one is interested in a particular set of actors or topics. For example, those super keen on local device regulation should look at Chapter 3, or those interested in AI agents might skip to Chapter 4. Skipping around like this will be easier for those who already have a broad knowledge of different actors in the technology ecosystem. Policymakers or civil society folk with this knowledge could skip from here to skim the chapters that look the most interesting to them, and then read the concluding chapter for our overall summary and cross-stack takeaways. The final section of each stack chapter discusses the most promising paths forward for better governance at that layer.

Some might find this book uncomfortable to read. Some of the governance methods we write about facilitate surveillance, consolidate the power of tech firms or threaten the openness of the Internet. "You'll give the bad guys ideas!" one colleague protested with despair upon hearing about our project. We hope that the reader will see that many of these ideas have long been in motion, that a lack of knowledge is no barrier to their progress and that they are not always heading in a positive direction. To shine a light on the AI stack allows us to discuss the trade-offs inherent in its governance. Such trade-offs by and large are made behind closed doors today. We hope to illustrate and thereby critique such modes of future governance, and show how, with more balance and moderation, they could be better crafted

in the future.

OUR THREE PREMISES

If you're sticking with us for the long haul, we'd like to tell you more about what we're trying to do in the pages that follow.

If society becomes interested in stopping bad (or at the very least troubled) people from doing bad (or at the very least deeply ambivalent and potentially quite problematic) things with AI tools, how should we approach that problem? Which issues and actors should we be thinking about? Let's tackle our three scoping arguments in turn.

ONE: WE NEED TO UNDERSTAND THE STACK

Much of the critical legal or social scientific work on AI systems begins by invoking the metaphor of the 'black box.' High-stakes algorithmic applications in the public or private sector are, following this logic, opaque, technically complex and difficult to understand. The implications are a set of challenges frequently summarized as fairness, accountability and transparency. How do we know if a system is biased or leading to unfair outcomes, and mitigate that if possible?[11] How do we observe a system, or try to get it to explain its decisions to us in a way that allows it to be held to account?[12]

This is not a wholly new type of problem when it comes to technology. Consider even a simple, well-defined task, such as getting people from A to B on an airplane. While airplanes are certainly designed to have the potential to safely get people from A to B, their practical safety and functionality relies on a maddeningly complex system of global certification, testing, auditing, training and schooling, regulation and accident investigations.[13] As a growing body of work on 'AI supply chains' has sought to illustrate, many popular AI systems are also extremely complex and distributed, without a single 'black box' sitting inside one company or actor that can be unpacked and analyzed.[14] Even where it might look like there is some discrete black box that can be handed over to a business to deploy, this is often a carefully maintained illusion.

Governance is emerging in these networks of actors, just as it has in relation to the Internet itself. Some self-organization is to be expected in private governance arrangements. It is unwise to attempt to regulate a complex network without knowledge of the empirical and

11: See generally Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness and Machine Learning: Limitations and Opportunities* (MIT Press 2023).

12: Lilian Edwards and Michael Veale, "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For" (2017) 16(1) Duke Law & Technology Review 18 DOI: 10/gdxthj.

13: Braden R Allenby and Daniel R Sarewitz, *The Techno-Human Condition* (MIT Press 2011).

14: When people inside a platform company are asked where the 'algorithm' behind a piece of functionality is, they often struggle to put their fingers on its exact location. The anthropologist Nick Seaver, who studied this phenomenon in a large music recommender company, noted that the "algorithm" was a "collective product, and consequently everyone felt like an outsider to it." See Nick Seaver, "Algorithms as Culture: Some Tactics for the Ethnography of Algorithmic Systems" (2017) 4(2) Big Data & Society 205395171773810 DOI: 10/gd8fdx. On AI supply chains, see e.g. Jennifer Cobbe, Michael Veale, and Jatinder Singh, "Understanding Accountability in Algorithmic Supply Chains" in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '23, ACM June 2023) DOI: 10.1145/3593013.3594073.

theoretical features of such environments.[15] When interlinked systems are involved, before you can govern them, you have to get to their beat, to understand their rhythm and anticipate the impact of disturbances you might try to introduce.[16]

Rather than just focusing on the activity of large developers such as OpenAI that have been at the center of much AI policy, we aim to weave together a larger picture of AI governance and coordination amid the multi-actor mess that we find. By doing so, we hope to illustrate a path forward for commentators and scholars that moves away from single-minded focuses on 'AI safety,' and tries to resonate with the context of how AI systems are being deployed and developed, rather than a highly simplified — and often misleading — 'black box' version of how they might be curtailed and constrained by a single actor.

There are at least five locations of current and potential governance that we need to think about.

The **cloud** (Chapter 2) provides the computing resources upon which so many actors in the generative AI pipeline rely. Researchers rely on the cloud to produce models. Companies use the cloud to fine-tune and deploy models. Cloud firms sell models directly and indirectly: mediating easy API access to both proprietary and open systems; selling value-added services like integration, easy user interfaces and 'safety filters'; renting out foundational computing resources like access to GPUs and virtual machines; and even serving as the fundamental hosting services to the rest of the intermediaries in the AI industry. The emerging regulatory focus on cloud computing centers on export controls and the ways to prevent purported bad actors from getting control over the powerful chips that could be used to develop strategically or 'existentially' dangerous models. Yet this kind of 'know your customer' logic is only scratching the surface of the emerging regulatory role of these providers, who are already becoming embroiled in the abusive activities that directly or indirectly rely on their services, and developing advanced techniques to monitor and disrupt them.

Advanced consumer-grade **local devices** (Chapter 3) are allowing users to run certain models on hardware they own. In doing this, people can deploy software in a way that escapes simpler centralized observability and makes policy intervention much more difficult. The way that hardware works has often been shaped in the past by key computational gatekeepers faced with difficult policy challenges. For example, many operating system producers like Apple, Google and Microsoft limit ripping or recording copyrighted multimedia through digital rights management (DRM) techniques. They limit access to certain on-device sensors which could be misused, only making them available to chosen applications on app stores. They limit users' ability to grant privileges to these sensors themselves to software that has not received their blessing. One can govern at the hardware layer, but how will the dynamics of usability versus control play out in a world of easily trained and traded open models that can fit snugly onto the new machine learning hardware being integrated into many common devices?

Network (Chapter 4) intermediaries are little discussed in AI governance but are important in the processes of both training and running models. Data for generative AI models is often scraped from Web services, which are in turn often protected against attacks by network services firms like Cloudflare. Such intermediaries are important in both development

15: Andrés Guadamuz, *Networks, Complexity and Internet Regulation: Scale-Free Law* (Edward Elgar Publishing 2011) 215.

16: Donella H Meadows and Diana Wright, *Thinking in Systems: A Primer* (Earthscan 2009) 171; Donella H Meadows, "Dancing With Systems" https://donellameadows.org/archives/dancing-with-systems/ accessed 2 September 2024.

— to shape the original datasets built — and at runtime. As AI services try to reduce 'hallucination' by grounding results in real-time Web searches, and AI 'agents' are used to interact with remote services on behalf of users, network intermediaries can block and exert other governance power on malicious actors hoping for limitless automated access.

Hosts (Chapter 5) are services where users can upload AI systems and AI-related datasets, which typically can be downloaded by others — and, depending on the business model, queried, tweaked or built upon by users around the world. Increasingly, these platforms are becoming important AI intermediaries, governing the visibility, ease of access and other dimensions of how individuals and organizations interface with generative AI models. Platforms like Hugging Face and Civitai have become extraordinarily popular in and impactful upon the open-source ecosystem, hosting tens of thousands of models since their inception. Mobile app stores, while having different overall functionality, are also increasingly hosting third-party AI apps, which clearly provide the easiest entry point for many people to use — and abuse — image generation models. These hosting platforms are emerging as important governors, and are slowly being targeted by governments and other stakeholders trying to get them to take down certain AI tools, although they are struggling to develop robust processes through which to do so.

Finally, popular **comms** (Chapter 6) intermediaries are gatekeepers sitting between users and generative AI material. Social networks, video and image sharing platforms, online forums and messaging providers are all places where those interested in model misuse create communities and share how such misuse can be carried out. These platforms have also been experimenting with detection techniques intended to find and flag AI-generated or modified media, but the poor performance of these systems has led instead to efforts to develop technical standards for content provenance. These efforts, however, are also limited in various ways and can be circumvented, so the platforms that feature the 'last mile delivery' of potentially malicious and/or illegal content created elsewhere have begun developing coordination mechanisms through which they might be able to share some information about certain content and certain actors with other stack actors, and thus also get more visibility into deeper dynamics in the misuse ecosystem.

These different locations of potential governance, as we will show, often blur and interact. Sometimes actors span multiple layers — Meta, for example, runs many user-generated content platforms while also uploading models to Hugging Face, running network infrastructure (e.g. WhatsApp) and producing consumer hardware. Stack governance locations are not the only places where real-world AI-related policy is happening or needed, but they are key chokepoints where the *functionality* of AI systems can be affected, and malicious uses potentially disrupted.

TWO: TECHNOREGULATORY REGIMES MATTER

Across all of these layers of the stack, we need to understand the varying mechanisms of control that are enabled by, facilitated by and constrained by technology. We have already called the salient methods of control in the generative AI ecosystem forms of technoregulation.

This can be formally and simply defined as using technology in order to enforce a rule.[17] Technoregulation — such as the filters that AI developers like OpenAI might develop to try to prevent ChatGPT or DALL·E from producing hateful content — is, unlike classic regulation, often pursued voluntarily by corporate actors, precisely because no legal authority is required to exercise control through technology. We use the term here to broadly refer to various forms of governance — the tools and practices through which various actors in the AI stack seek to shape the way that the tools they build are used — of varying types, from informal norms to explicit policies and the technical systems used to enforce them.

Technoregulation has several practical benefits over classical regulation, as well as its own set of concerns. It avoids the thorny problem of which jurisdiction misdeeds occur in when they happen on the Internet.[18] It brings an enforcement scale, speed and automaticity that rival the technologies that are embroiled in the mischief itself.[19] Yet the ability of such technoregulation to curtail human freedoms and 'configure' those humans according to the whims of private actors has also led to calls to take its impact upon the world more seriously.[20] For example, should technoregulation be limited to preventing only the worst types of misbehavior, in order to avoid the moral damage possible in a society where the ability to misbehave is removed?[21]

Technoregulation often involves a multitude of different interlinked actors and a network of governance relationships and stakeholders that coax it into being. Governance scholars often use the idea of a 'regime,' or 'regime complex,' to describe such a layered, multi-actor ecosystem of regulation on a specific policy issue area.[22] In the AI ecosystem, the technoregulatory governance regime involves existing as well as new legal frameworks explicitly being developed to regulate AI, the realities of industry practice and self-regulatory guidelines and policies, the recommendations and non-binding propositions made by various civil society groups and multi-stakeholder bodies increasingly active on the topic of generative AI, as well as new codes of conduct, standards and technical tools for harm mitigation that might be more or less prominent depending on the specific AI system in question.

17: We use the term technoregulation following Roger Brownsword, "What the World Needs Now: Techno-Regulation, Human Rights and Human Dignity" in Roger Brownsword (ed), *Global Governance and the Quest for Justice - Volume IV : Human Rights* (Hart Publishing 2004) DOI: 10.5040/9781472563217 , but the rest of this note lists sibling terms; Karen Yeung, "Towards an Understanding of Regulation by Design" in Roger Brownsword and Karen Yeung (eds), *Regulating Technologies: Legal Futures, Regulatory Frames and Technological Fixes* (Hart 2008) (design-based regulation); Lawrence Lessig, *Code: Version 2.0* (2. ed., Basic Books 2006) (code as a regulatory modality); Joel R Reidenberg, "Lex Informatica: The Formulation of Information Policy Rules through Technology" (1998) 76(3) Texas Law Review 553, (information policy rules, or lex informatica); Mireille Hildebrandt, "Legal and Technological Normativity: More (and Less) than Twin Sisters" (2008) 12(3) Techné: Research in Philosophy and Technology 169 DOI: 10.5840/techne20081232, (looser form of technoregulation which includes rules that dissuade rather than prevent).

18: Reidenberg (n 17).

19: ibid.

20: 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 ch 7.

21: Roger Brownsword, "Neither East Nor West, Is Mid-West Best?" (2006) 3(1) SCRIPTed: A Journal of Law, Technology and Society 15; Roger Brownsword, "Code, Control, and Choice: Why East Is East and West Is West" (2005) 25(1) Legal Studies 1 DOI: 10.1111/j.1748-121X.2005.tb00268.x; Roger Brownsword, "Law, Liberty, and Technology" in Roger Brownsword, Eloise Scotford, and Karen Yeung (eds), *The Oxford Handbook of Law, Regulation and Technology* (Oxford University Press 2017) DOI: 10.1093/oxfordhb/9780199680832.013.2.

22: Kenneth W Abbott, "The Transnational Regime Complex for Climate Change" (2012) 30(4) Environment and Planning C: Government and Policy 571; Julia Black, "Constructing and Contesting Legitimacy and Accountability in Polycentric Regulatory Regimes" (2008) 2(2) Regulation & Governance 137.

Crucial to our understanding of technoregulation is the chokepoint.[23] This is a classic concept for Internet and platform regulation, but has been largely ignored in the literature on machine learning and society. As a long tradition in technology law has shown, the nature of the Internet provides countless ways for people to conceal their identities when breaking the law or upsetting powerful actors like copyright holders. For that reason, historically, harmful, malicious, undesirable or illegal activity online has been typically disrupted by, or through, *intermediaries* that can be pressured or regulated to change their policies or their behavior.[24]

Various actors in the AI stack have come to realize that *they* are potential governance chokepoints and are starting to encounter questions as to the extent of their legal and moral responsibilities, and how they might act upon them. We explore many such realizations in detail throughout this book. Providers of models behind APIs implement ever more complex filtering to detect certain types of illegal or policy-violating content. Model hosting services allowing AI systems to be downloaded are trying to create policies and processes to decide the conditions under which a model is removed from the platform, as regulators around the world increasingly become interested in dual-use models and the intermediaries involved in allowing people to access them.[25] Laws are emerging requiring AI firms to watermark outputs, and social media platforms to act — in ways often yet to be defined — on such watermarks.[26]

While this kind of work on networked governance in the context of technology policy has a very long history, perhaps its most recent manifestation has been in the interdisciplinary literature on content moderation, platform governance and 'trust and safety.'[27] This literature describes an increasingly complex multi-layered system of regulation shaping how major Internet providers with various business models deal with potentially harmful and illegal content that is shared on those platforms by third-party users. That work has not only shown how important informal governance and the private regulation of large technology companies have become for today's digital policy ecosystem, but has also helped illustrate some of its failures, shortcomings and knock-on effects. There are complex politics at play, with more than a decade of actors exploiting, or weaponizing, interdependencies in the platform governance ecosystem for certain policy outcomes.[28] Examples include payment processors shaping content moderation rules for adult content websites,[29] app stores being pressured to remove social networks with limited forms of content moderation,[30] ISPs following the recommendations of

23: Jack L Goldsmith and Tim Wu, *Who Controls the Internet? Illusions of a Borderless World* (Oxford University Press 2006); Natasha Tusikov, *Chokepoints: Global Private Regulation on the Internet* (University of California Press 2016).

24: Goldsmith and Wu (n 23).

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

26: See Chapters 2 and 6.

27: Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018); Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech" (2017) 131(6) Harvard Law Review 1598; Sarah T Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale University Press 2019); Robert Gorwa, *The Politics of Platform Regulation: How Governments Shape Online Content Moderation* (Oxford University Press 2024).

28: Robert Gorwa, "What Is Platform Governance?" (2019) 22(6) Information, Communication & Society 854 DOI: 10.1080/1369118X.2019.1573914 ; José Van Dijck, Tim de Winkel, and Mirko Tobias Schäfer, "Deplatformization and the Governance of the Platform Ecosystem" (2023) 25(12) New Media & Society 3438 DOI: 10.1177/14614448211045662.

29: Zahra Stardust and others, "High Risk Hustling: Payment Processors, Sexual Proxies, and Discrimination by Design" (2023) 26(1) CUNY Law Review 57; Emily Pollak, "The Internet Is for Porn... or Is It? Fair Access to Financial Services and the Need for OnlyPorn Legislation" (2022) 17(1) Brooklyn Journal of Corporate, Financial & Commercial Law 191.

30: Josh Cowls and Jessica Morley, "App Store Governance: The Implications and Limitations of Duopolistic Dominance"

private bodies to filter specific content without any express legislative basis,[31] or DDoS attack protection being removed from controversial websites by Internet infrastructure providers.[32]

As we argue in this book, the generative AI ecosystem will also inevitably be governed through this kind of public-private technoregulation. The history of technoregulation, however, is one of imperfect exercises in governance — messy, unreliable and with plenty of accompanying side-effects upon market structures and the viability of less centralized networked technologies. It offers a crucial vocabulary for understanding both the possibilities and the limits of the rapidly emerging policy area that is of interest to us here.

THREE: THE RISKS ARE ALREADY HERE

Finally, this book is about the misuse of generative AI models. Unlike much work on AI policy, especially the work produced in the small yet increasingly influential community of 'AI safety' research, we do not focus on extreme or existential risks to humanity. 'AI safety' researchers are concerned about the misuse of actual AI systems, but have so far, for a mixture of philosophical, ideological and economic reasons, focused on what they argue are 'extreme risks,' arguably to the exclusion of many important others. When considering the governance of these risks, this school of thought has focused on potential institutional mechanisms that could monitor for, and react to, the sudden emergence of world-threatening AI systems, to "maximize the chances of the world safely navigating this transition."[33] Such literature gives little focus to the specific actors and infrastructures involved in either enabling these models or limiting their capabilities, largely because it deals in an unspecified future and draws few connections to or from what we know about governing networked technologies over recent decades. Some might even argue that powerful machines and models make such previous experience irrelevant. Yet the history of Internet and technology law shows intermediaries coming back again and again as a core site of governance. There is still quite a bit of gas left in that particular regulatory tank.

Long-term questions are not completely irrelevant, but are nonetheless disconnected from the core questions facing industry about how their models are used and misused today. Centering the fantastical at the expense of the inevitable is a risky strategy too. Models that exist at the time of writing, far from displaying any type of 'general' or 'super' intelligence, are already being used by a range of actors to supercharge various forms of problematic, illicit and illegal activity. These kinds of shorter-term concerns are of huge import, yet have largely

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.

31: TJ McIntyre, "Child Abuse Images and Cleanfeeds: Assessing Internet Blocking Systems" in Ian Brown (ed), *Research Handbook on Governance of the Internet* (Edward Elgar Publishing 2013) DOI: 10.4337/9781849805025.00020; TJ McIntyre, "Internet Censorship in the United Kingdom: National Schemes and European Norms" in Lilian Edwards (ed), *Law, Policy, and the Internet* (Hart Publishing 2019).

32: Suzanne van Geuns and Corinne Cath-Speth, "How Hate Speech Reveals the Invisible Politics of Internet Infrastructure" (*Brookings*, 20 August 2020) https://perma.cc/DEX7-5MTA accessed 10 October 2021; Electronic Frontier Foundation, "ISPs Should Not Police Online Speech—No Matter How Awful It Is." (*Electronic Frontier Foundation*, 29 August 2023) https://www.eff.org/deeplinks/2023/08/isps-should-not-police-online-speech-no-matter-how-awful-it accessed 2 August 2024, chapter 1.

33: Allan Dafoe, "AI Governance: A Research Agenda" (27 August 2018) https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf accessed 4 August 2024, 6.

fallen by the wayside in today's interdisciplinary conversations on AI governance.

Interestingly, before the latest generative AI hype cycle, the large industry players acknowledged the importance of malicious use openly. For example, OpenAI cited 'generating fake news content, impersonating others in email, or automating abusive social media content production' as the three particular types of misuse motivating their decision to 'stage' the release of the weights of an earlier language model, GPT-2.[34] Google, similarly, did not release their 280-billion-parameter 'Gopher' model in 2021, long before ChatGPT's release, citing similar public safety considerations.[35]

More recent publications from large AI firms instead downplay the immediate risks of AI misuse and seek to redefine 'dangerous capabilities' to focus on the potential for 'autonomy' of AI 'agents.' This has resulted in Google defining 'dangerous capabilities' to specifically exclude the "highly contextualized misuse scenarios" such as scamming or generating sexual abuse imagery, and instead testing for capabilities such as 'self-reasoning,' 'self-proliferation' and 'persuasion or deception.'[36] There is nothing inherently wrong with testing for concerns raised by those analyzing the risks of 'artificial general intelligence' (AGI), no matter how far-fetched and implausible these concerns may actually be, but frameworks that by design exclude dangerous capabilities that are known to be present when humans use these tools maliciously, and include only those which are speculative and rely on the agent itself to design and execute an entire malicious workflow, are naïve at best and conveniently dishonest at worst. One could even speculate that these issues are omitted because they seem too hard for the AI developers to mitigate. While critics have rightly identified a certain PR energy to all these purported concerns — 'criti-hype' that serves to both scare and animate[37] — there is nonetheless a lesson to be learned from this relatively recent change of emphasis.

The earlier and in many ways comparatively richer work from OpenAI draws some useful distinctions, and further highlights some specific omissions from the views of Google and many more recent efforts. In 2019, OpenAI constructed three idealized threat actors: low-skilled, limited-resource actors; actors with moderate programming skills able and willing to build malicious products; and advanced persistent threats with organized groups and resources. A large part of their stated logic to release their full GPT-2 model was, they claimed, the lack of evidence of these low- and moderate-resource actors mobilizing to make usable tools, embedded in the AI ecosystem, that facilitated misuse.[38]

As we'll explore in the ensuing pages, follow-up studies would surely not find such a scarcity

34: Irene Solaiman and others, "Release Strategies and the Social Impacts of Language Models" (*arXiv*, 12 November 2019) DOI: 10.48550/arXiv.1908.09203.

35: Laura Weidinger and others, "Taxonomy of Risks Posed by Language Models" in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22, ACM June 2022) DOI: 10.1145/3531146.3533088 ; Jack W Rae and others, "Scaling Language Models: Methods, Analysis & Insights from Training Gopher" (*arXiv*, 21 January 2022) DOI: 10.48550/arXiv.2112.11446.

36: Mary Phuong and others, "Evaluating Frontier Models for Dangerous Capabilities" (*arXiv*, 5 April 2024) DOI: 10.48550/arXiv.2403.13793.

37: Inioluwa Deborah Raji and others, "The Fallacy of AI Functionality" in *2022 ACM Conference on Fairness, Accountability, and Transparency* (2022) DOI: 10.1145/3531146.3533158 . See specifically Lipton who claims that OpenAI's staged release strategy exists "to catapult immature work into the public view, and [play up the] human safety aspects of work that doesn't yet have intellectual legs to stand on." Zachary C Lipton, "OpenAI Trains Language Model, Mass Hysteria Ensues" (*Approximately Correct*, 17 February 2019) https://www.approximatelycorrect.com/2019/02/17/openai-trains-language-model-mass-hysteria-ensues/ accessed 13 August 2024.

38: Solaiman and others (n 34).

of tools today, nor a paucity of low- and moderate-resource malicious actors. In this book, we focus on these kinds of harms — medium term, medium risk. They are not extreme in the sense of catastrophe — no AI-generated killer viruses, global cyberphysical meltdown or hell-bent 'superintelligent' robots sacrificing all human life at the altar of their church of paperclips.[39] Instead, they are the kinds of harms that are already here, and that industry is actively dealing with: the emerging practices of abuse monitoring, platform governance and 'trust and safety' that already are having huge implications for generative AI systems and how we use them.

Crucially, this book is also not about governing every controversial use of AI systems. A range of serious harms already accompany AI systems working exactly as intended by actors developing or using them. Some might consider this itself misuse, particularly if certain AI systems are unethical or illegal — issues of social scoring,[40] biometric surveillance,[41] digitally mediated segregation and redlining,[42] ubiquitous online tracking,[43] military applications,[44] and exploitative price-gouging of consumers.[45] Furthermore, AI systems come with serious pollution-like 'externalities,' and are often optimized for certain tasks at the expense of the broader environment and the individuals, communities and social fabrics that constitute it.[46] For generative systems, the theft of creative and artistic labor,[47] the working conditions of data workers and content moderators,[48] the loss of dignity in decisions about individuals,[49] potential damage to the formation of the self,[50] the threats to perceptions of authentic or trustworthy information or multimedia,[51] the extraction of labor and reshaping of the world

39: This is a reference to an argument made in Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press 2014).

40: Oscar H Gandy, *The Panoptic Sort: A Political Economy of Personal Information* (Westview Press 1993); Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin's Press 2017).

41: Kashmir Hill, *Your Face Belongs to Us: A Secretive Startup's Quest to End Privacy as We Know It* (Random House 2023).

42: Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code* (Polity 2019).

43: Michael Veale and Frederik Zuiderveen Borgesius, "Adtech and Real-Time Bidding under European Data Protection Law" (2022) 23(2) German Law Journal 226 DOI: 10.1017/glj.2022.18.

44: Elke Schwarz, *Death Machines: The Ethics of Violent Technologies* (Manchester University Press 2018); Denise Garcia, *The AI Military Race: Common Good Governance in the Age of Artificial Intelligence* (Oxford University Press 2023) DOI: 10.1093/oso/9780192864604.001.0001.

45: Natali Helberger and others, "Digital Fairness for Consumers" (*BEUC, the European Consumer Organisation*, Brussels, Belgium, March 2024) https://perma.cc/Z46G-AVP2 accessed 5 October 2024; Frederik Zuiderveen Borgesius, "Price Discrimination, Algorithmic Decision-Making, and European Non-Discrimination Law" (2020) 31(3) European Business Law Review 401.

46: Bogdan Kulynych and others, "POTs: Protective Optimization Technologies" in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAT* '20: Conference on Fairness, Accountability, and Transparency, ACM 2020) DOI: 10/ggjf5m.

47: Trystan S Goetze, "AI Art Is Theft: Labour, Extraction, and Exploitation: Or, On the Dangers of Stochastic Pollocks" in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '24, Association for Computing Machinery 5 June 2024) DOI: 10.1145/3630106.3658898.

48: Roberts (n 27); James Muldoon, Mark Graham, and Callum Cant, *Feeding the Machine: The Hidden Human Labour Powering AI* (Canongate Books 2024).

49: Meg Leta Jones, "The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood" (2017) 47(2) Social Studies of Science 216 DOI: 10/f93mxz.

50: Mireille Hildebrandt, "Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning" (2019) 20(1) Theoretical Inquiries in Law 83 DOI: 10.1515/til-2019-0004; Sylvie Delacroix and Michael Veale, "Smart Technologies and Our Sense of Self: Going Beyond Epistemic Counter-Profiling" in Mireille Hildebrandt and Kieron O'Hara (eds), *Life and the Law in the Era of Data-Driven Agency* (Edward Elgar 2020) DOI: 10/gfzvz9; Julie E Cohen, "What Privacy Is For" (2012) 126 Harvard Law Review 1904.

51: Robert Chesney and Danielle Keats Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National

of work,[52] and the environmental cost of training and querying have all been posed as crucial questions for AI policy.[53]

Our focus on a subset of relatively unambiguous malicious misuse has an analytical purpose rather than a normative one. We are not making a claim that preventing organized crime from running financial scams is more important than addressing wide-scale discrimination embedded into generative machine learning systems. Our claim is instead that focusing on harms that are more normatively settled within law and societal institutions (as regrettably many of the serious topics, such as the technological extent of indirect discrimination or 'disparate impact' above, are not) allows us to focus on mechanisms of prevention and disruption, and the distribution of responsibility. As a result, we keep the definition of malicious misuse loose. It is not tied to illegality in any specific jurisdiction, although elements of the activities we discuss are typically already illegal in many places. We use the term to pose a fairly simple question. If democratic societies were to speak in one voice to say that a certain practice should no longer exist, how could and would that rule be enacted in practice? Put another way: there will be governance stakeholders (civil society, academics, journalists, citizen groups) that will try to change AI technologies in an effort to curb misuse. How could they actually do it?

The sharp analytic focus on mechanisms of disruption that our scope affords also helps us shed light on how networks of actors in the AI stack might coordinate to limit all types of illegitimate deployment, or reduce or mitigate the externalities of their systems, generative or not. Intermediaries do play a role in many of these broader efforts to shape AI policy — from Google employees organizing in response to the firm's involvement in military AI through *Project Maven*,[54] to cloud firms like IBM changing position on their support of biometric analysis.[55] But we think that significant research is being undertaken to unpack the many potential impacts that AI is having on labor, culture and the environment.[56] The angle we take in this book seeks to complement this work by enhancing our understandings of intervention points and governance strategies in the generative AI ecosystem, without dupli-

Security" (2019) 107 California Law Review 1753 DOI: 10.2139/ssrn.3213954.

52: Hamid R Ekbia and Bonnie A Nardi, *Heteromation, and Other Stories of Computing and Capitalism* (MIT Press 2017).

53: Sasha Luccioni, Yacine Jernite, and Emma Strubell, "Power Hungry Processing: Watts Driving the Cost of AI Deployment?" in *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '24: The 2024 ACM Conference on Fairness, Accountability, and Transparency, ACM 3 June 2024) DOI: 10.1145/3630106.3658542; Emily M Bender and others, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability and Transparency in Computing Systems (FAccT 2021)* (ACM Conference on Fairness, Accountability and Transparency in Computing Systems (FAccT 2021), 2021) DOI: 10/gh677h.

54: Thao Phan and others, "Economies of Virtue: The Circulation of 'Ethics' in Big Tech" (2022) 31(1) Science as Culture 121 DOI: 10.1080/09505431.2021.1990875; Jathan Sadowski, "'Open Secrets': An Interview with Meredith Whittaker" in Thao Phan and others (eds), *Economies of Virtue – The Circulation of 'Ethics' in AI* (Institute of Network Cultures, Amsterdam 2022).

55: Mark Wilding, "IBM Promised to Back off Facial Recognition — Then It Signed a $69.8 Million Contract to Provide It" (*The Verge*, 31 August 2023) https://www.theverge.com/2023/8/31/23852955/ibm-uk-government-contract-biometric-facial-recognition accessed 5 October 2024.

56: Similarly, why do we look at the particular parts of the generative AI stack that we have chosen? Other places in the broader computing supply chain are crucial to consider when reducing externalities, but have less direct connection to the topics at the heart of this book. Some, such as data-labeling labor, raw material extraction and energy generation, are huge social issues in and of themselves, but are nonetheless disconnected from downstream abuse and misuse and the disruption of them. A model can be abused whether or not data labelers feeding into it had fair wages and psychological support. It can be abused whether it was trained using datacenters powered by solar or by methane. While the social and environmental costs of the mining operations for rare earths used in the GPUs used to train a system are undoubtedly important, these costs don't transform the functionality of the system. Governing the actors at the far ends of these global networks is important, but not for the reason of preventing downstream AI misuse.

cating or providing redundant, shallow summaries of policy issues dealt with in much more depth by others elsewhere.

OUR MAIN TOOLS AND ARGUMENTS

So, if you're following us so far, a quick recap. Our book focuses on what might be done to stop malicious, AI-enabled activity across the whole 'stack' of different technology actors, many of whom are not very frequently considered in AI governance conversations. Our understanding of misuse is medium-risk, medium-term — we're interested in the kinds of immediate harms that are already present in the current state of the art in genAI development (and also latent in many already existing models far from the frontier). As a kind anonymous reviewer of this book noted, this makes it "not a sweeping book about AI but a book about a particular subset of problems confronting AI governance today."

As indicated above, our argument moves through varied conceptual layers, with one per chapter: the 'cloud' layer (providers such as Amazon, Google Cloud, OpenAI or Anthropic); the 'local' layer (on a user's device, which might be a phone or computer); the 'network' layer (where services being accessed by AI systems, such as those being scraped, try to decide who can or cannot use them); the 'hosting' layer (focusing on where open-source models can be downloaded, with firms such as Hugging Face and GitHub center stage); and finally the 'comms' layer (social media and similar, where victims are reached and perpetrators often share tutorials with each other). At each of these layers, we share stories, mechanisms of governance, interlinks with other layers, and the consequences and wisdom of seeking different levels of intervention at this layer.

As we go through this book, we will lay out our theoretical tools and the necessary prerequisite knowledge for the ensuing technical material step-by-step. The notes contain extensive citations, references and, at times, further exposition for the curious. We'll begin by introducing a key conceptual tool that recurs across the chapters, and by previewing a few themes and findings that have come out of our investigation before we start our journey through the AI stack.

HOOKS AND LOOPS

Our book develops a new conceptual tool for understanding distributed, technologically mediated regulation: governance *hooks* and governance *loops*. Because actors at a particular layer of the AI stack have only limited visibility into and agency over the kinds of third-party and customer activity that they may wish to disrupt, they might not be able to put the brakes on that activity on their own. Loops are types of interventions they could choose instead. Loops are often forms of technoregulation, not in and of themselves designed to prohibit certain forms of action, but rather designed to be somewhat persistent across the stack and permit future forms of action.[57] We can think of them as a form of governance residue.

57: This is linked to the concept of 'affordances' in human–computer interaction, possibilities for action readily perceivable by an actor. This concept comes however from the field of design, relating to what individuals think they can do with things, whereas we are more interested in designers enabling certain possibilities and communicating the potential use in a structured,

Loops can be baked into the function of certain AI systems or platformized infrastructures, sitting dormant until they are called upon when required. A governance *hook* is a practice that then mobilizes a pre-existing loop, permitting new types of information gathering, analysis and perhaps also action. Hooks hook into loops. These hooks and loops form a core part of our policy playbook, and we think of them a bit like bits of Velcro — a material itself made from hooks and loops that is quite powerful when both parts are connected. These kinds of interventions involve new ways of thinking about distributed technoregulation that can help disparate actors coordinate to meaningfully govern some types of behavior in a maximally effective manner.

We look at an array of hooks and loops in this book. Loops include watermarking, cloud usage logging, software notarization, secure enclaves, provenance metadata, *robots.txt* preference files, behavioral use licenses and 'model cards.' Hooks often connect to these to block, filter or otherwise treat software and generated content differently on the basis of where they find the loops and what they reveal. Behind these hooks might be a single actor, or a more coordinated effort. They might be a public actor or a private actor. In any case, our model of hooks and loops attempts to shine a light on how technoregulation works when it is distributed, rather than the current approach to its analysis, where it is typically thought of as concentrated in the hands of a single actor.

DIGGING DEEP

Hardware matters for AI governance, but not only in the ways you may have been led to believe. In the ensuing pages, we show the various ways in which robust, difficult-to-work-around interventions designed to prevent AI misuse are increasingly being deployed — or would need to be deployed — deep in the AI stack, built into hardware so that systems can potentially monitor what is being done on users' devices to, for example, attest that it is genuine or follows a set of rules, or disrupt it if it is malicious. But in contrast to the typical conversation around global AI governance, the key players here are not chipmakers and GPU providers; they're typically operating system developers that have the highest level of access to the consumer-grade hardware that most AI misuse is being conducted with. GPUs are extraordinarily badly placed to assess whether the use they are supporting is good or bad; instead, the hardware divisions of firms like Apple, Google and Microsoft will be key actors in the transnational fight against AI misuse. This is not without cost. Their involvement brings with it a range of thorny issues, from power concentration to privacy violations.

FRICTION

Our book will discuss a range of different forms of AI misuse, and touch on a wide variety of systems with different functionality that nonetheless are commonly lumped together under the generative AI moniker. Readers will be disappointed to learn that constantly thwarting

formal manner. The term affordances is not about the formalization and communication of potential actions that can be taken by other (governance) actors, as was recognized by its popularizer in his later work. See Donald A Norman, *The Design of Everyday Things* (2nd edn, Basic Books 2013) (introducing the concept of 'signifiers' to try to fill this gap). This language is unwieldy for our purpose in any case, and 'affordances' as a term confuses many audiences, so we do not use it here.

all prospectively problematic uses of these manifold different tools is, unfortunately, given the nature of the globally distributed Internet and the devices that run on it, inherently impossible.

But thoughtful interventions in the AI stack can insert friction, and make it more difficult for certain users to do certain kinds of things. Friction-oriented strategies can minimize the costs of clumsy technoregulation, by instead targeting low-hanging fruit and trying to deter less motivated and less expert users. For instance, while there are inherent difficulties to eliminating or even durably reducing the ability of open (and even closed-source, cloud-hosted) image generation models to produce sexual forgeries, there are many comparatively easy interventions that could be taken to combat the availability of mobile apps that provide such deepfaking capabilities as a service. The availability of guidance, helpers and tutorials; the advertising and publicity of malicious tools; the slowing down of certain types of services to make them harder to scale; the requirement to circumvent technical defenses that may not be perfect. Applying friction is a creative act of discovering where it can be best located, and throughout the book, we identify a wide range of areas that may have some promise for those wishing to limit the misuse happening with AI systems.

SEGMENTATION

When taking governance at any stack layer to its limits — not always a great or desirable idea for society, as we will discuss — we find that such moves may frequently create undesirable or even dangerous knock-on effects. This means that deep, durable, effective and difficult-to-circumvent interventions against AI misuse by AI intermediaries will require what we call 'segmentation': deploying techniques that allow the actors being governed to be separated out such that different modalities of governance can be applied across different groups.

Segmentation allows different governance logics to apply to different groups. If you manage to first segment out bots from humans, you can require different methods of authentication and assessment of trustworthiness and intent for each, rather than trying to fit them all into the same mould. The same applies for regulating devices: perhaps researchers and consumers have different needs? The former can more easily have organizational and due dilligence requirements placed upon them than the latter. Segmentation is necessary for nuanced governance, and we see it appear throughout this work.

BAD BAD NOT GOOD

The story told in the ensuing pages, as we travel, layer by layer, through the generative AI 'stack,' is in one sense an old tale about disrupting illegality online. In another, it's a new and rapidly emerging story about computational, infrastructural and institutional complexity. This book is our effort to understand the difficult governance puzzles that come with a technology that as easily creates arresting imaginary cat photos as warzone disinformation; that can produce not only handy stock footage but also disturbing, highly targeted sexual abuse videos; and that can generate summaries of research papers with a few clicks and in parallel spit out long, convincing, tailored scam messages pretending to be from your uncle who apparently is stuck at the airport without his bank card.

Part of our provocation in writing this book is an effort to reframe the terms of the policy debate, and to bring things down to their nitty-gritty technical reality. It can be tempting to get caught up in the highfalutin' language of international declarations and policy debates about AI as artificial 'intelligence,' which inherently is accompanied by a feeling of inevitability: 'intelligence' connotes something slippery and raw, that will constantly, naturally progress on its own. But once you peel back the curtain on many AI systems — how they function, how they are used, and how their uses can be shaped and modified — one sees that these systems are little more than complex assemblages of software. We like to think of them as 'bad computers' — bad in a moral sense, as some of the systems we will discuss have been specifically built with evil in mind, but also bad in an admonishing sense, because these systems, while certainly not magical or superintelligent, are in fact very brittle, opaque, complex and tricky to govern.

Policymakers and the public are about to be hit by the next wave of bad computers, and need to know how to respond. How should they do so? In the following pages, we will discuss how, looking across the stack, we can identify possibilities to minimize and manage at least some of the most malign uses of these technologies in our collective future. We will also explore the broader consequences of those potential governance strategies — for the future of the Internet, computation and society.