Chapter 6

Comms: Distributing Tutorials and AI Output

Michael Veale and Robert Gorwa

Chapter 6 of *Malice, Models and Middlemen: The Full Stack Fight Against AI Misuse*, pages 147–175

Abstract

The Internet is all about communication, and every element of the stack covered within this book so far is involved in transmitting information to and through machines. But some parts of the Internet have been explicitly designed for humans to get in contact with other humans: messaging apps, social networks, video and file-sharing services. Let's call them 'comms platforms.'

Comms platforms are somewhat distinct in the AI stack in other ways. Perhaps more than any other layer, they recur in multiple distinct roles throughout the entire pipeline of model misuse. In this chapter, we begin by looking at how comms platforms help animate and coordinate malicious activity by hosting tutorials, communities of potential and actual abusers, as well as ads and marketplaces for services used by these abusers. We then examine how these platforms are used to intermediate between AI generation and the victims of various types of malice. This might be through more private means, such as direct messages sent by a scammer, or more public means, such as the circulation of non-consensual sexual forgeries. Although this initially may seem like just a content moderation problem — something which has been written about with some depth when it comes to AI-generated media — we explore the various constraints that can make this a much more difficult problem than often acknowledged in the extant policy conversation. Throughout, we focus on the new issues raised by multipurpose generative AI systems: how might these comms platforms respond?

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 6

https://homepages.ucl.ac.uk/~ucqnmve/malicemodelsmiddlemen/chapter/6/

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.

6 COMMS: DISTRIBUTING TUTORIALS AND AI OUTPUT

The Internet is all about communication, and every element of the stack covered within this book so far is involved in transmitting information to and through machines. But some parts of the Internet have been explicitly designed for humans to get in contact with other humans: messaging apps, social networks, video and file-sharing services. Let's call them 'comms platforms.'

Comms platforms are somewhat distinct in the AI stack in other ways. Perhaps more than any other layer, they recur in multiple distinct roles throughout the entire pipeline of model misuse. In this chapter, we begin by looking at how comms platforms help animate and coordinate malicious activity by hosting tutorials, communities of potential and actual abusers, as well as ads and marketplaces for services used by these abusers. We then examine how these platforms are used to intermediate between AI generation and the victims of various types of malice. This might be through more private means, such as direct messages sent by a scammer, or more public means, such as the circulation of non-consensual sexual forgeries. Although this initially may seem like just a content moderation problem — something which has been written about with some depth when it comes to AI-generated media — we explore the various constraints that can make this a much more difficult problem than often acknowledged in the extant policy conversation. Throughout, we focus on the new issues raised by multipurpose generative AI systems: how might these comms platforms respond?

ILLICIT INSTRUCTION

Sometime in 2023, a person sitting somewhere booted up their computer and loaded the Tor browser, the U.S. government-funded software enabling access to the 'darkweb.' They navigated to a URL ending in *.onion*, the domain ending for 'hidden services' only available through the Tor network designed to robustly hide the location of the server. On this hidden service, a small community of English speakers discussed and shared child abuse imagery. The person posted the following:

> *Good day lads, I am here to encourage more of you to try to make your own artificial intelligence child girlies [...] Look into Automatic1111's webui github project, this will allow you to run stable diffusion models locally on your GPU.*[1]

Research indicates such guidance on how to use AI for child abuse material generation on the darkweb is increasingly common. These forums feature detailed threads on topics like

1: Deanna Davy and Samantha Lundrigan, "Artificial Intelligence-Produced Child Sexual Abuse Material: Insights from Dark Web Forum Posts" (*International Policing and Public Protection Research Institute, Anglia Ruskin University*, July 2024) https://www.aru.ac.uk/-/media/Files/ipppri/IPPPRI-Insight-No-1-AI-CSAM.pdf accessed 13 August 2024, 7. Academic researchers access this type of content using special scraping services designed to return text but never images.

"Photo Resources for AI and Deepfaking Specific Girls," with experienced users sharing tips on the best models, training data and technical setups.[2]

Where instruction manuals are hosted on the darkweb, there are few intermediaries in the stack that can locate or take them down. For example, to locate and take down the servers of the illicit darkweb marketplace *The Silk Road*, known for narcotics and hitmen, security experts believe the U.S. National Security Agency had to tap and monitor undersea cables off the coast of Iceland. This gambit worked, as the infamous darknet site's admin login page was set up without proper encryption over the clearweb.[3] Such interception capabilities were so secret, and the NSA's domestic surveillance of U.S. citizens like Silk Road founder Ross Ulbricht so legally shaky, that the FBI is alleged to have fabricated a 'parallel construction' of plausible alternative detection methods in the hope of not having their sketchily collected evidence struck out.[4] For our purposes, this story illustrates the extremes that governments must go to in order to disrupt such illegal services, and that they can't necessarily rely on traditional Internet intermediaries to help them. By the time someone starts hosting content on the darkweb, unless they seriously slip up in their operational security, normal methods of governance through intermediaries or domestic law enforcement powers are often rendered ineffective.[5]

Yet many groups providing such information or selling services operate much less deeply buried than the darkweb and its hidden services. A wide range of communities provide guides to those looking to use AI to generate non-consensual intimate imagery of, for example, celebrities and influencers. Telegram's unencrypted group messaging is home to several of these communities.[6] Telegram is also home to those who sell illicit AI tooling to criminal groups, as we showed in Chapter 3. Chapter 2 also showed that the tools and platform around hacking Azure credentials for image generation credits to create non-consensual intimate imagery were hosted on the anonymous 'pastebin' service, Rentry — another sort of comms platform where anyone can upload information to serve to others. But many AI tutorials are easily available on some of the Internet's most popular sites, like YouTube. As we will see, dealing with these tutorials can be difficult even for the best-resourced platforms.

FOR EDUCATIONAL PURPOSES ONLY

Format Boy is a scamfluencer. His extremely active Telegram channel, with nearly 20,000 subscribers, regularly sees him advertising software clearly well-suited for scams, showing re-

2: Internet Watch Foundation, "How AI Is Being Abused to Create Child Sexual Abuse Imagery" (*IWF*, October 2023) https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf accessed 15 April 2025, p. 23.

3: Robert Graham, "Reading the Silk Road Configuration" (*Errata Security*, 3 October 2014) https://blog.erratasec.com/2014/10/reading-silk-road-configuration.html accessed 7 March 2026.

4: U.S. law enforcement alleged that a leaky CAPTCHA sent the real IP address of the hidden service outside the Tor network, which observers have noted is technically implausible given the service's specific set-up. On this practice and the concerning reasons behind it, as well as this specific Silk Road example, see Natasha Babazedeh, "Concealing Evidence: Parallel Construction, Federal Investigations, and the Constitution" (2018) 22(1) Virginia Journal of Law & Technology 1; Graham (n 3).

5: Pascal Tippe and Adrian Tippe, "Onion Services in the Wild: A Study of Deanonymization Attacks" (2024) 2024(4) Proceedings on Privacy Enhancing Technologies 291 DOI: 10.56553/popets-2024-0117, showing that the vast majority of Tor deanonymization in U.S. court cases can be attributed to such slip-ups.

6: Emanuel Maiberg and Samantha Cole, "AI-Generated Taylor Swift Porn Went Viral on Twitter. Here's How It Got There" (*404 Media*, 25 January 2024) https://www.404media.co/ai-generated-taylor-swift-porn-twitter/ accessed 27 May 2025.

alistic fake calls where individuals are transformed into a fully suited Elon Musk, a typical impostor target. Format Boy posts pictures of stacks of hundred-dollar bills, and promises followers wealth if his instructions are followed.[7] Format Boy draws on the aesthetics and subculture of the Yahoo Boys, young men who scam online and are frequently depicted in Nigerian film and music. To some, the Yahoo Boys are get-rich-quick Robin Hoods: deploying all the necessary tactics to survive, and possibly even thrive, in a difficult context with staggering youth unemployment. For others, they have damaged Nigeria's international reputation and had real consequences for the country's attractiveness as a place to do business, just as the '419' or 'Nigerian Prince' scammers did before them.[8] A popular YouTube channel claiming to be Format Boy lists videos on voice swapping, face-changing, image editing and how to compromise the protections on smartphones to place phone calls with altered media inputs (Figure 6.1). Each video, and the channel as a whole, features the disclaimer *for educational purposes only*.

It is unclear to what extent Format Boy's teachings are themselves facilitating scams, or whether his real victims are the wannabe scam artists, coughing up real dollars for software and information that reeks of snake oil. Many other influencers around the world are also trying to profit from the AI boom in this way.[9] Nonetheless, Format Boy's YouTube accounts have been repeatedly shut down, resetting his follower count every time, and encouraging his move to less moderated platforms such as Telegram.[10]

Moderating such videos poses complex questions for a platform like YouTube. Many major user-generated content platforms already limit the sharing of some kinds of instructional speech. But they also do not clearly communicate how potentially dual-use information will be handled. YouTube, a platform full of tutorial videos, unsurprisingly has perhaps the most extensive policies in this area, but even these are shown to be vague on detailed examination.

YouTube's policies typically have some explicit intent requirement, meaning that moderators need to decide whether content is 'for educational purposes only.' The most relevant prohibitions forbid videos "[d]emonstrating how to use computers or information technology with the *intent* to [..] cause serious harm to others."[11] Some further policies highlight other forms of prohibited content, including videos that provide instructions on how to forge official documents, purchase drugs on the darkweb, optimize eating disorders or effectively commit suicide. In all areas, exemptions to the policy can be argued for on educational, documentary, scientific and artistic grounds.

Yet hosting video or other forms of tutorial online is trivial. Even if nuanced policies were applied to major platforms like YouTube, the long tail of middle platforms before the darkweb would likely continue to host it, keeping it up given its ambiguity, and benefiting from the

7: Some coverage of Format Boy can be found here Matt Burgess, "How to Win Followers and Scamfluence People" (*Wired*, 19 May 2025) https://www.wired.com/story/format-boy-yahoo-boys-scam-influencers/ accessed 7 March 2026. The authors looked through his Telegram feed in March 2026.

8: See generally Carlos Barragán, "The Romance Scammer on My Sofa" (*The Atavist Magazine*, 24 June 2023) https://magazine.atavist.com/2023/the-romance-scammer-on-my-sofa-nigeria-yahoo-boys accessed 7 March 2026.

9: Jason Koebler, "Where Facebook's AI Slop Comes From" (*404 Media*, 6 August 2024) https://www.404media.co/where-facebooks-ai-slop-comes-from/ accessed 8 March 2026.

10: Obidike Okafor, "Who Is Format Boy?" (*Businessday NG*, 7 June 2025) https://businessday.ng/news/article/who-is-format-boy/ accessed 8 March 2026.

11: YouTube, "Community Guidelines" (*Google*, January 2026) https://github.com/OpenTermsArchive/pga-versions/blob/main/YouTube/Community%20Guidelines.md (emphasis added).

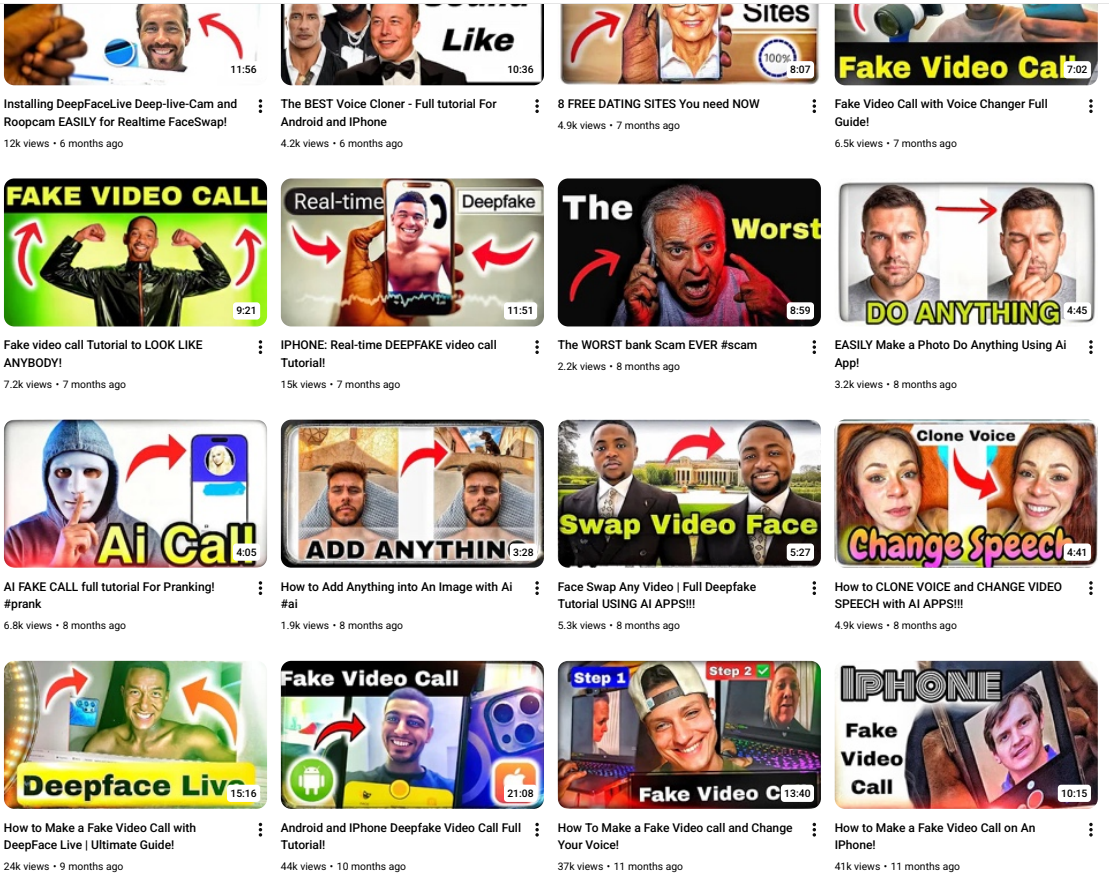

Figure 6.1: YouTube tutorials from popular user @Formatboy1234 capable of supporting misuse, presented as 'for educational purposes only.' It is unclear whether this account is a popular copycat or maintained by the owner of the eponymous Telegram channel; previous accounts have been reportedly shut down. Retrieved by authors, March 2026.

clicks that come from almost — but not quite — breaching policies.[12] Voluntary moderation is quickly stressed when communities are deeply determined to seek out instructional content and will move to where it is.

Going beyond measures taken voluntarily by platform governors here would involve re-examining the role that law should play in the regulation of instructional speech. This is far from a new issue. It is, perhaps surprisingly, one of the oldest political and legal debates in Internet regulation.

12: Paul Lewis, "'Fiction Is Outperforming Reality': How YouTube's Algorithm Distorts Truth" *The Guardian* (2 February 2018) https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth accessed 8 March 2026; Tarleton Gillespie, "Do Not Recommend? Reduction as a Form of Content Moderation" (2022) 8(3) Social Media + Society 20563051221117552 DOI: 10.1177/20563051221117552 .

MAYHEM MANUALS

In April 1995, Timothy McVeigh bombed a federal government complex in Oklahoma City, killing 168 people and injuring 684 more. Days later, the *New York Times* reported that human rights groups had, for weeks before the bombing, observed a 'surge' in bomb-making instructions on online bulletin boards, including of the same fuel and fertilizer combination used by McVeigh in Oklahoma. Four days after the bombing, a post offered to provide "complete details of the bomb used in Oklahoma City, and how it was used and could have been better" to "anyone who properly requests it."[13]

McVeigh's bombing really had nothing to do with the Internet — while he had a teenage dalliance with hacking, by the time of his attack, he lived a technologically disconnected life, obsessed with firearms and finding community on the traveling American gun show circuit.[14] Yet the incident fanned the flames of a long-burning panic about the availability of 'mayhem manuals' online. Transcribed 'textfiles' of print manuals such as *The Anarchist Cookbook* were indeed being uploaded to bulletin boards such as The Temple of the Screaming Electron (TOTSE), often by kids.[15] Better understood as angsty rebels rather than terrorists, teenagers hacked insecure telco infrastructure to get to these bulletin boards, 'phreaking' and playing dialtones down the line to fool providers into connecting long distance calls for free, avoiding huge phone bills and angry parents.[16] Sometimes, there were tragic consequences: in 1988, four Maryland teenagers — reportedly the best chemistry students in their school — were killed when a bomb they were secretly building in one of their family's garages exploded.[17] The students, as well as more than a thousand other individuals, were subscribers to a bulletin board called 'Pyromaniac Production Systems,' which "provided information on car bombs, rockets, pipe bombs, and land mines" to hobbyists interested in understanding how they worked.[18]

After Oklahoma, Congress held a hearing which would shape the legal debate around these kinds of guides for the next decade. Senator Herb Kohl set the scene:

> *Indeed, in just an hour-and-a-half of surfing the Internet, my staff and the Congressional Research Service uncovered the following materials: "How to Build an Atomic Bomb," "The Terrorist Handbook," "How to Pick a Lock," "The World Sex Guide to Prostitution" and many, many sex bulletin boards. Under no circumstances should this be available to our children. Period.*[19]

The risks of children having access to atomic bomb creation instructions aside, there was a

13: Keith Schneider, "Terror in Oklahoma: The Internet; Talk on Bombs Surges On Computer Network" (*The New York Times*, 27 April 1995) https://www.nytimes.com/1995/04/27/us/terror-in-oklahoma-the-internet-talk-on-bombs-surges-on-computer-network.html accessed 21 January 2026, n.p.

14: Jeffrey Toobin, *Homegrown: Timothy McVeigh and the Rise of Right-Wing Extremism* (Simon and Schuster 2024).

15: Jason Scott (director), Episode 6: HPAC (Hacking Phreaking Anarchy Cracking), BBS: The Documentary (2005) https://commons.wikimedia.org/wiki/File:BBS_The_Documentary,_Episode_6.webm accessed 21 January 2026.

16: ibid.

17: Ann Larabee, *The Wrong Hands: Popular Weapons Manuals and Their Historic Challenges to a Democratic Society* (Oxford University Press 2015) 139.

18: ibid 139.

19: *The Availability of Bomb-Making Information on the Internet: Hearing Before the Subcomm. on Terrorism, Tech., & Gov't Info. of the S. Comm. on the Judiciary*, 104th Cong. (1995).

growing demand to regulate in this area. Legislators emphasized how easy it was to download instructions to make explosives with easily accessible materials and tools, although they may have made the matter worse. Then-Senator Joe Biden read several pages of detailed instructions on how to produce thermite and how to extract gunpowder from shotgun shells into the Congressional Record, where it is easily found to this day.[20] Biden, together with Senator Dianne Feinstein, passed an initial proposal to prohibit bomb-making tutorials, although it was almost immediately amended out in a subsequent law due to constitutional concerns, and replaced with a requirement that the Department of Justice commission a report on the issue.[21]

Both news and popular media had, even before the attack, been preoccupied with amateur bombings informed by mail-order books published by paramilitary publishers such as Paladin Press, or disseminated through online bulletin board services. Just like Format Boy in 2026, advertisements for such information in the 1990s were keen to point out that these bomb-making instructions were, of course, 'solely for informative purposes.'[22]

In the United States, there has been much discussion (and little conclusion) on exactly when 'instructional,' 'informational' or 'crime-facilitating' speech should be considered protected under the First Amendment, and when individuals may instead be held liable for it.[23] The main Supreme Court case here, *Brandenburg v. Ohio*, focuses on speech that might persuade someone to commit an offense, requiring demonstration that such advocacy 'is directed to inciting or producing imminent lawless action and is likely to incite or produce such action.'[24] Tutorials would typically struggle to pass the hurdle of proving such immediate risk, and additionally are often teaching things that could be, even in theory, used for positive or negative ends, including detection or defense. The Supreme Court has however left plenty of wiggle room around how the law should or could work here. Refusing to take a later case, Justice Stevens emphasized that the *Brandenburg* holding did not explicitly encompass instructional speech. What test should apply in this area is therefore still unknown.[25]

Later cases from respected lower courts have muddied the waters further. In one lower court, a publisher was successfully sued for selling an instruction guide entitled *Hit Man: A Technical Manual for Independent Contractors* that did, indeed, lead to a contract killing.[26] On recommendation from the book, the hitman had bought a specific gun, erased its bore markings with a file, drilled out the serial numbers, built a silencer and shot two victims in the eyes from a distance of six to eight feet.[27] One of the strangest things about this situation was that the book itself was written by a divorced mother of two who, having had a novel rejected by publishers, instead turned it into this 'guide' by pulling together insights from "books, television, movies, newspapers, police officers, [her] karate instructor, and a good

20: 142 Cong. Rec. S3454–S3455 (daily ed. Apr. 17, 1996).

21: Larabee (n 17) 147.

22: Schneider (n 13) n.p.

23: See discussions in Eugene Volokh, "Crime-Facilitating Speech" (2004) 57(4) Stanford Law Review 1095; Martin H Redish, *The Logic of Persecution: Free Expression and the McCarthy Era* (Stanford University Press 2005); Samantha Mitchell, "First Amendment Speech Protections in A Post-Dobbs World: Providing Instruction on Instructional Speech" (2023) 91(4) Fordham Law Review 1521.

24: *Brandenburg v. Ohio*, 395 U.S. 444, 447 (1969).

25: *Stewart v. McCoy*, 537 U.S. 993 (2002) (Stevens, J., respecting the denial of certiorari).

26: *Rice v. Paladin Enterprises, Inc.*, 128 F.3d 233 (4th Cir. 1997), cert. denied, 523 U.S. 1074 (1998).

27: Larabee (n 17) 98.

friend who is an attorney."[28] Far from being a hitwoman, she did not even own a gun.

In one further case, the U.S. Supreme Court permitted restrictions on even peaceful speech and training as a form of 'material support' delivered in a manner 'coordinated' with a specific terrorist organization, as it had a 'legitimating' effect — even though such restrictions triggered the highest standard of constitutional review that is, in almost every case, fatal to a speech-restrictive statute.[29] Given that influencers and tutorial communities online do often have commercial connections with their audiences, such as through subscriptions or software sales, this does open a potential avenue to legal rules that prohibit somewhat direct tutoring connections. Yet this has only added to what has been described as a 'state of almost total confusion' in the courts, with the scholarship too 'in almost as confused a state.'[30]

The U.S. also prohibits some forms of teaching or providing information on how to make or use explosives, destructive devices or weapons of mass destruction, on the basis of rules eventually passed following Biden and Feinstein's interventions above.[31] There are two prongs to this offense — someone must distribute this information with the intent to further a violent federal crime, or they must distribute this information to someone they know has such an intent themselves. Having barely been used in almost three decades, this provision was recently found constitutional by the Fourth Circuit in December 2025.[32] That case however focused on the perhaps easier case of a speaker knowing the intent of a specific listener, shining little light on the more difficult question of broadcast, such as online posts and tutorials.

Some forms of instructional content, including through broadcast, are more clearly prohibited in other jurisdictions. The UK, known for its draconian and wide-ranging speech restrictions on expression relating to serious crime and national security issues, may have gone the furthest of modern democracies. UK law already makes it a serious offense to possess, collect or make a record of 'information of a kind likely to be useful to a person committing or preparing an act of terrorism,' with no regard needed as to whether the accused intends or is connected to, even in the remotest sense, any act of terrorism at all.[33] Possession of a "paedophile manual" — training or guidance for offenses committed against real children — is also an offense in most of the United Kingdom.[34] A similar offense exists in Germany.[35] Neither offense covers AI-generated images, but the UK Parliament passed an amendment in 2026 to extend the 'manual' offense to instructions to create synthetic child abuse imagery.[36] The terrorist information offense and the non-AI 'paedophile manual' offense are already 'priority illegal content' for all platforms, big and small, to proactively remove under the UK's Online

28: David Montgomery, "If Books Could Kill" (*The Washington Post*, 26 July 1998) https://www.washingtonpost.com/wp-srv/style/features/paladin.htm accessed 18 January 2026, n.p.

29: *Holder v. Humanitarian Law Project*, 561 U.S. 1 (2010) (as only the second time in history a First Amendment restriction has passed the 'strict scrutiny' test.)

30: Redish (n 23) 106.

31: 18 U.S.C. §842(p).

32: *United States v. Arthur*, No. 24-4306 (4th Cir. Dec. 3, 2025).

33: Terrorism Act 2000, s 58; For criticism, see Jacqueline Hodgson and Victor Tadros, "How to Make a Terrorist Out of Nothing" (2009) 72(6) The Modern Law Review 984 DOI: 10.1111/j.1468-2230.2009.00777.x.

34: Serious Crime Act 2015, s 69. Note that this offense extends to England, Wales and Northern Ireland; no equivalent offense currently exists in Scotland.

35: Strafgesetzbuch [German Criminal Code] (StGB) §176e.

36: Crime and Policing Act 2026 s 75 (uncommenced as of August 2026).

Safety Act.[37] It seems inconceivable that the Act would not be amended by the Secretary of State using delegated authority to include the new AI manual offense in the near future.[38]

These issues are back on the table in the 2020s, especially as easily available large language models can produce these kinds of instructions, or even act as interactive guides or textbooks. As discussed throughout this volume, closed models can, with the help of some clever prompts, have their safety protections 'jailbroken,'[39] while open-weight, downloadable models can be queried without many of these protections at all. Indeed, 'abliterated' models — models that have had their tendencies to refuse prompts erased — have very few protections at all.[40] The flexibility of these models in answering questions about tutorials or phenomena might even make them more dangerous than the quite easily located bomb-making 'textfiles' of old. Matthew Livelsberger, who injured seven people and took his own life in Las Vegas in early 2025 using a Tesla Cybertruck rigged with explosives, built the bomb and learned how to trigger it with interactive help from ChatGPT.[41] Just months later, a fertility clinic in Palm Springs was bombed, injuring four people. The main suspect in that case also had a series of AI chatbot records regarding bomb-making.[42]

Furthermore, these types of offenses are being stretched further by new types of AI misuse that are, in many ways, much easier to commit. Despite the near-obsession of the largest AI firms with framing safety around bioweapons and 'biorisk,' research has argued that the main constraint preventing such weapons from being built is rarely a lack of *secret knowledge*.[43] To find realistic risks that might be amplified, we might want to look at areas with no difficult supply chains of chemicals, reagents, lab equipment or materials to deal with. Some malice just requires a computer. Consider all the activities we have talked about in this book so far — scams, child abuse material, non-consensual imagery and more. Models serve as tutorials not just to mayhem in the real world, but to mayhem in the virtual world, instructing potential wrongdoers in model misuse and abuse.

While instructional speech might be removed from platforms, how this should be done will remain fraught. While jurisdictions like the UK seem to have no qualms with passing rules on limiting such speech, and applying them in turn to platforms, the U.S. legal landscape is considerably more uncertain. Even if slapping 'educational use only' disclaimers on content is not quite enough to avoid takedown by either law or private community standards, tutorial creators can take a step back into generality to ensure that the instructions are 'dual-use enough' to surf the edge of policy acceptability. Furthermore, we have to note that LLMs' own tutorial capabilities are also considerable, particularly when they are 'uncensored.' How should platforms reason about how-to videos on how to run uncensored LLMs — something which feasibly may have security research uses or be desired by users seeking unfiltered results

37: Online Safety Act 2023, sch 5(1)(n), sch 6(8).

38: ibid s 222.

39: See Chapter 2.

40: See Chapters 3, 5.

41: Kashmir Hill, "When Chatbots Are Used to Plan Violence, Is There a Duty to Warn?" (*The New York Times*, 26 February 2026) https://www.nytimes.com/2026/02/26/technology/chatbots-duty-warn-police.html accessed 6 March 2026.

42: Department of Justice (United States), "Washington State Man Arrested on Federal Charges Alleging He Provided Material Support to Palm Springs Fertility Clinic Bomber" (4 June 2025) https://www.justice.gov/opa/pr/washington-state-man-arrested-federal-charges-alleging-he-provided-material-support-palm accessed 6 March 2026.

43: Aidan Peppin and others, "The Reality of AI and Biorisk" in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '25, Association for Computing Machinery 23 June 2025) DOI: 10.1145/3715275.3732048.

in a context of government-managed information control — if those tutorials thereby become the most facilitative of future misuse?[44]

LAST MILE DELIVERY

Comms intermediaries don't just host and spread tutorials, how-to guides and coordination mechanisms for actors using these AI systems. They have a significant role in facilitating illegal content dissemination when it occurs.

Think about how the unwanted consequences of many malicious uses of AI often manifest when generated content is distributed to individuals or the public at large. Discerning whether a deepfake is a legitimate satire of a billionaire, or a targeted campaign to defraud would-be investors, can often only be achieved via reference to the specific context of its dissemination — often through a comms platform. Put another way, determining whether dual-use AI is being used for good or bad will often hinge on the context of its communicated outputs. Yet the nature of generative AI, as well as the way that many comms platforms function, means that their ability to detect AI-generated content and reason about it to make a good governance decision is enormously limited. Let's look at an example scenario before diving into the details.

PROMISES AND PITTFALLS

In late 2023, a French interior designer in her early fifties was on a ski holiday. With the encouragement of her daughter, she downloaded Instagram for the first time, and publicly posted some family photos from the trip. All proceeded according to the ordinary logic familiar to those who have used social media in the past — a new profile, slowly growing a following by finding old friends and acquaintances, with a few random follows sprinkled in — until she received an unusual direct message from someone claiming to be connected to a famous Hollywood actor, who ostensibly stumbled upon her profile and wanted to speak with her. Eventually, our protagonist was put directly in touch with a special new pen pal: Brad Pitt.[45]

Virtual Brad Pitt was kind, and nice to talk to. He asked the Frenchwoman often about her work and troubles, in marked contrast to her ex-husband. He was a romantic, often sharing touching poems and music. Sometimes, he sent over images of his extremely handsome self. Could this really be Brad Pitt that she was talking to? It seemed highly unlikely, but it wouldn't be the craziest thing to have happened in this extremely online world. And the Frenchwoman wasn't born yesterday: she knew how to conduct a reverse-image search. The selfies appeared to be new, legit pics that didn't exist online elsewhere.

44: The boundary between code and instructional speech causes significant legal headaches, see generally Andrea M Matwyshyn, "Hacking Speech: Informational Speech and the First Amendment Constitutional Law" (2012) 107(2) Northwestern University Law Review 795.

45: Zoé Lauwereys, "Ils ont usé de mon humanité: Anne, victime d'un faux Brad Pitt, raconte l'engrenage de l'arnaque" (*Le Parisien*, 15 January 2025) https://www.leparisien.fr/faits-divers/ils-ont-use-de-mon-humanite-anne-victime-dun-faux-brad-pitt-raconte-lengrenage-de-larnaque-15-01-2025-YRQAFIY3Z5AUNG5CNCPGGKKXLY.php accessed 27 May 2025.

Eventually, Brad — surprisingly a fluent French speaker — opened up about his deep debts and constant financial struggles, perhaps unusual for someone with a publicly reported net worth of close to half a billion dollars. As the Frenchwoman had just received a significant sum as part of divorce proceedings, she was happy to help her favorite actor and new long-distance boyfriend get back on his feet. As Virtual Brad made excuses as to why they couldn't yet meet in person, he assured her that their relationship was already making waves in Hollywood: the woman received a high-quality video in which a news program talked about Pitt's 'exclusive relationship with one special individual,' mentioning the Frenchwoman by name.

At some point in their months-long correspondence, the LA tabloids broke the news that offline Brad Pitt had begun dating the jewelry designer Ines de Ramon. Confronting Virtual Brad about his duplicity, and unswayed by his excuses, the Frenchwoman broke off contact. But the emotional and economic damage had been done. The woman was reportedly scammed out of more than 800,000 euros.[46] Her wealth and gullibility were widely mocked by the French public when her (pseudonymous) story was broken by a major broadcaster.[47]

These kinds of romance scams are becoming more prevalent, and various AI tools are helping make them more deceptive and effective. As we have discussed in this book, AI supports the production of synthetic content, real-time fake audio and video, and even the automation of chats, likes, posts and more.[48] So many of these interactions happen via comms platforms, both social networks and instant messengers. Meta Platforms' services alone accounted for over 54% of scams involving deceptive bank transfers in 2024.[49] Should Instagram have detected that the victim was being contacted by a sketchy account? Could the messenger through which Virtual Brad was communicating with the Frenchwoman have analyzed the media files that Brad sent to her and predicted the likelihood that AI had generated or edited them? Should the platform have tipped off the user? Blocked the sender? Reported them to law enforcement? What, if anything, should comms platforms do to prevent scams, and are there AI-specific aspects to this challenge that they should or could make use of in doing so?

CONTENT-ORIENTED DETECTION

If platforms could reliably detect AI-generated content, they could act upon that information in various ways. But the detection of AI content is neither easy nor reliable, and comes with a range of social and technological trade-offs and considerations that are crucial to consider. Comms platforms may attempt to do content-oriented detection through a variety of techniques, including by conducting predictive classification of images on-device or in the cloud, deploying older tricks from media forensics or by facilitating the adoption of watermarking schemes. We'll discuss these — and their limits — in turn, before moving to an overview of the additional strategies that try to detect the use of AI by looking beyond the

46: Maxime Poul, "'J'aimais l'homme avec qui je conversais:' une femme se fait soutirer 830 000 euros par un faux Brad Pitt" (*Le Parisien*, 13 January 2025) https://perma.cc/W77Q-QUBP accessed 27 May 2025.

47: Laura Gozzi, "French Woman Duped by AI Brad Pitt Faces Mockery Online" (*BBC News*, 15 January 2025) https://www.bbc.com/news/articles/ckgnz8rw1xgo accessed 27 May 2025.

48: See Chapters 3, 4.

49: Revolut, "Consumer Security and Financial Crime Report H2'24" (31 March 2025) https://cdn.revolut.com/pdf/Revolut_Consumer_Security_Report_H2_2024.pdf accessed 8 March 2026.

content itself.

APPS AND OPERATING SYSTEMS

We have talked extensively in this book about mobile operating systems in the context of querying models (Chapters 3, 4). They recur again here in a different context — as comms intermediaries involved when individuals send or receive messages. Apps and operating systems have both been under growing pressure to act to prevent AI-driven misuse.

Chinese law has, since 2022, placed strong emphasis on the development and deployment of anti-fraud technologies by telecommunications businesses, such as mobile phone operating systems.[50] European law has been moving in a similar direction, allowing payment providers to transfer liability for fraud to comms platforms, under certain circumstances, where the fraud originated on their platforms and they did not take sufficient action to prevent it.[51] This is likely to incentivize the co-operation of such providers, and the development of anti-fraud technologies, similar to the requirements in Chinese law.

Honor, one of the most popular smartphone manufacturers in China, launched a tool built into the operating system for the identification of AI-generated video calls, using an on-device convolutional neural network trained on a large number of fake and real images, regularly updated through security patches.[52] It later deployed another tool for the detection of synthetic voices. Both Huawei's 'Star Shield' and Google's 'Scam Detection' tools claim to be able to detect common scam scripts in audio calls through on-device transcription and classification by AI models.[53] If this worked, it might have helped tip off the Frenchwoman to be more careful about her new texting partner — although she may have simply dismissed the warning — or tip off the messaging app about a common risk with that sender — although such profiling comes with privacy risks and risks of false positives.

When anti-fraud technology runs at the operating system level, it typically works primarily on-device. Companies put great emphasis on this, typically in terms of privacy and security.[54] Privacy is especially important as such tools break the rules that usually apply to other apps, such as being permitted to use restricted APIs (discussed in Chapter 3) to, for example, view every pixel on the screen at any given time.

There are fundamental engineering reasons other than privacy for this to happen on device, however. Scanning cannot happen in the cloud if messaging services are end-to-end encrypted. If users' devices automatically uploaded audio or video they receive, it would create high cloud compute costs for providers, use individuals' data allowances and poten-

50: 中华人民共和国反电信网络诈骗法 (Anti-Telecommunications and Internet Fraud Law of the People's Republic of China) of 2022, arts 30, 32, 35.

51: The political agreement between the Parliament and Council on the Third Payment Services Directive (PSD3) and new Payment Services Regulation (PSR) reportedly provides for this; however, the final, agreed text has not been published at the time of writing, and so the described requirement is only an approximation from this reporting.

52: Honor, "MagicOS 10.0: Security Technical White Paper" (*Honor Device Co Ltd*, 15 October 2025) https://perma.cc/4CQB-7894 accessed 13 January 2026.

53: Huawei, "隐私 (Privacy)" (*Huawei Developers (HarmonyOS)*, 24 November 2025) https://perma.cc/SU2W-QXUR accessed 13 January 2026; Lyubov Farafonova and Alberto Nieto Pastor, "New AI-Powered Scam Detection Features to Help Protect You on Android" (*Google Security Blog*, 4 March 2025) https://perma.cc/B4S9-FT3K accessed 13 January 2026.

54: Honor, "AI 隐私安全白皮书 (AI Privacy and Security White Paper)" (*Honor Device Co Ltd*, October 2025) https://perma.cc/CY3X-R4ZX.

tially strain the network as a whole. Putting such tools on-device may, however, drain users' batteries, particularly on older phones. In any case, in China, analysis of traffic associated with such anti-fraud tools indicates that not everything *is* happening on device, with some metadata (such as who-communicates-with-whom, or who has which apps installed) sent to central authorities.[55]

Indeed, the empirical question of how effective these tools are in actually boosting user awareness and disrupting scams remains unclear. It is easy to see how the same non-tech-savvy users, overwhelmed by the features already on their smartphone, might also be convinced that any alert or notification is something they should ignore. While these systems claim high accuracy in their marketing materials, typically of 95% or above, Chinese reporters testing these protections found that none of the phones that were tested detected actual AI faceswaps, even with the tools easily available to the journalists.[56]

UNDETECTED

The failure of these kinds of detectors isn't some failure specific to Chinese smartphone providers. 'AI-generated' and 'AI-modified' are capacious categories which can refer to a huge range of different types of media produced through various tools and techniques. Many AI systems claiming to do AI-powered content analysis are classic 'AI snake oil,' producing results that are little better than random despite their claims of stellar performance.[57] There are two reasons for this.

As one review of the many 'deepfake detectors' on the market put it, the most common strategy for AI content detection involves training a classifier to predict the likelihood that a piece of media is AI-generated after having fed it many instances of content labeled as 'AI' or 'not AI.'[58] This appears to be the approach followed by Honor and other Chinese device manufacturers. While this kind of tool can indeed achieve good performance on the labeled media, meta-reviews of the literature show that this approach cannot yet consistently generalize to media that the classifier has not seen before.[59] Research also indicates that advertised performance of models drops drastically when tested on realistic, real-world data.[60] After all, there are now so many forms of AI manipulation and generation — different models with different architectures that leave different types of visible or invisible features in media for AI systems to potentially detect — that it is very difficult for tools to make conclusive predictions on all kinds of media in a setting where no other information about the method of content

55: Open Technology Fund's Red Team Lab, "China's National Anti-Fraud Center - Security Assessment" (*Open Technology Fund*, 21 September 2022) https://www.opentech.fund/news/chinas-national-anti-fraud-center-security-assessment/ accessed 8 March 2026.

56: Yuting Zhang, ""AI 换脸"5 秒可检出? 实测三款手机未识别-南方都市报 · 奥一网 (Can "AI Face Swap" Be Detected in 5 Seconds? Tests Show Three Phones Failed to Identify It)" (*Southern Metropolis Daily*, 14 May 2025) https://perma.cc/AG63-C32X accessed 13 January 2026.

57: Arvind Narayanan and Sayash Kapoor, *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference* (Princeton University Press 2024).

58: Alistair Knott and others, "AI Content Detection in the Emerging Information Ecosystem: New Obligations for Media and Tech Companies" (2024) 26(4) Ethics and Information Technology DOI: 10.1007/s10676-024-09795-1 accessed 18 July 2025.

59: ibid.

60: Nuria Alina Chandra and others, "Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024" (*arXiv*, 27 May 2025) DOI: 10.48550/arXiv.2503.02857.

generation is known. Even a small false positive error rate, multiplied by the vast amount of shared content today, massively limits the realistic action that can be taken on flagged content.

The second problem is that AI detection in the real world, outside of controlled research settings, is adversarial. There are now many peer-reviewed papers in top machine learning conferences that have developed various strategies for identifying outputs from certain specific models — for example, identifying the distinctive statistical noise that is in the pixel matrices of images extruded by models like Stable Diffusion[61] — and many of these can be quite effective when deployed narrowly to try to find outputs from those models. But savvy operators can simply fine-tune their model using known detection techniques to change the outputs accordingly and render those detection methods ineffective.[62] This kind of *ex post* detection of AI content, whether deployed in the cloud by a large platform like Instagram or by a messaging provider on device (as in the Honor example mentioned previously), creates an arms race. The widespread public adoption of 'AI detection' tools simply incentivizes savvy actors to train and tweak models designed to evade those methods. Motivated attackers could even consider contaminating confirmed, real photos with AI generation artifacts to further sow confusion and skepticism about 'AI-generated' labels. It's difficult to imagine a future where AI content detectors can do more than provide a signal to be contextually interpreted, and actually become a reliably accurate tool to inform the governance of malicious genAI outputs downstream.

MEDIA FORENSICS

Less snake-oil-tinged than these software solutions are efforts rooted in more classic, physics-based methods for media forensics, which attempt to compare the characteristics of content to the constraints that must exist in the real world. For example, photo forensics looks at concrete aspects of an image, including how light interacts with the physical world, how it travels through different lenses, how it gets captured by cameras' sensors and how it gets transformed into digital files.[63] Common classic signs of forgery include inconsistent lighting, impossible vanishing points and shadows or the lack of a pattern of noise resembling a known camera. In more recent years, in response to fully generated images, physiological impossibilities have also become important to analysts. These have included inconsistent reflections in generated eyes or glasses, the interaction of light on the skin and even the rhythm of a pulse beneath it. Similar efforts have been deployed for other media too: audio forensics considers factors such as the impact the recording location will have on reverberation to understand inconsistencies in the waveforms.[64]

The problem with these techniques is that they don't easily scale. Manual forensic techniques require deep judgment and expertise, and cannot obviously be deployed automatically at the scale of billions of uploads per day.[65] Similarly to automated detectors, they also cannot

61: Zhendong Wang and others, "DIRE for Diffusion-Generated Image Detection" (*arXiv*, 16 March 2023) DOI: 10.48550/arXiv.2303.09295 accessed 9 March 2024.

62: Knott and others (n 58).

63: Hany Farid, *Photo Forensics* (MIT Press 2016) 3.

64: Robert C Maher, *Principles of Forensic Audio Analysis* (Springer 2018) DOI: 10.1007/978-3-319-99453-6 ch 5.

65: Hany Farid, *Fake Photos* (MIT Press 2019).

be conclusive on their own. It may not usually be possible to definitively 'prove' that an image is edited or manipulated, but one can assemble a range of evidence of abnormalities that might provide a strong indicator that an image isn't what it claims to be.

To help with the scale problem, there have been some efforts to build these kinds of strategies into easy-to-use detection pipelines. For instance, researchers at the State University of New York at Albany got ample media coverage in the late 2010s for their use of computer vision methods to analyze the eye movements in videos depicting people, with an attempt to predict the likelihood of it being AI-modified or generated by looking at unnatural blinking patterns.[66] Another tool, created by Intel researchers, claimed that many AI-generated videos featured visible and unusual blood vessel movements in the face that could be detected with their software.[67] We have very little knowledge about how well future AI outputs from leading models will pass these tests, yet there are good reasons to believe that many of the same problems facing predictive detection tools for previously unseen content will apply.

Major comms platforms could benefit tremendously from reliable automated identification of AI-generated content. They could choose to limit its reach until certain conditions had been met, filter it into specialized reviewing pipelines or generally help their trust and safety teams navigate what many have predicted will be an eventual deluge of problematic AI-generated material.[68] Yet there is little indication they are acting on the outputs of such tools. One reason may be an unwillingness to enter into the politically charged area of misinformation and disinformation often associated with fake content.[69] The kinds of harms and misuse we have been talking about in this volume are not subject to the same political contention as generic misinformation, and there is broad consensus that scamming and non-consensual intimate image dissemination, for example, should be disrupted somehow. Another reason may be platforms' unwillingness to subject synthetic content to different standards than 'authentic content.'[70] But detecting whether somebody is an impostor of another real person — a policy which may well apply evenly whether AI is involved or not — is a much harder test than detecting whether an image or piece of media is synthetically generated. AI content does not necessarily have to be treated differently by platforms for AI detection to be useful for them.

PROVENANCE DETECTORS

Another strategy that can be used by comms platforms is to use the watermarks that may be embedded in the outputs of certain AI models to help them detect if content is AI-generated

66: Yuezun Li, Ming-Ching Chang, and Siwei Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking" (*arXiv*, 11 June 2018) DOI: 10.48550/arXiv.1806.02877 accessed 18 July 2025; Tackhyun Jung, Sangwon Kim, and Keecheon Kim, "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern" (2020) 8 IEEE Access 83144 DOI: 10.1109/ACCESS.2020.2988660 accessed 18 July 2025.

67: Umur Aybars Çiftçi, İlke Demir, and Lijun Yin, "Deepfake Source Detection in a Heart Beat" (2024) 40(4) The Visual Computer 2733 DOI: 10.1007/s00371-023-02981-0 accessed 14 November 2025.

68: Samuel Woolley, *The Reality Game: A Gripping Investigation into Deepfake Videos, the next Wave of Fake News and What It Means for Democracy* (Hachette 2020); Gillespie, "Do Not Recommend?" (n 12).

69: Robert Chesney and Danielle Keats Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security" (2019) 107 California Law Review 1753 DOI: 10.2139/ssrn.3213954; Christopher T Marsden, Trisha Meyer, and Ian Brown, "Platform Values and Democratic Elections: How Can the Law Regulate Digital Disinformation?" (2020) 36 Computer Law & Security Review 105373 DOI: 10.1016/j.clsr.2019.105373.

70: Sarah A Fisher, Jeffrey W Howard, and Beatriz Kira, "Moderating Synthetic Content: The Challenge of Generative AI" (2024) 37(4) Philosophy & Technology 133 DOI: 10.1007/s13347-024-00818-9.

or not. As outlined in Chapter 2, some AI model providers, under pressure from many jurisdictions to add watermarks or secure metadata to their generated content, have relied on two strategies thus far: clipping 'secure metadata' onto media files to indicate their provenance, or embedding various identifiable watermarks into the generated media itself. There are also some attempts to blend the two, such as by attaching the metadata within the watermark.

While providers have faced watermarking and metadata obligations, the story regarding obligations on comms platforms is a messier one. Nevertheless, some jurisdictions have passed rules to more clearly steer the situation.

Detection laws

The EU's Digital Services Act requires large platforms to address systemic risks to the European Union, which may include disinformation, but it does not prescribe the manner in which this must be done, or oblige or even recommend watermark detection as a tool to achieve this.[71] There is some scope for this to be elaborated upon through, for example, codes of practice, but there is no explicit regime here to hook into the governance loop left by the watermarking obligation on AI providers in the EU AI Act. In Europe, it is unclear whether preemptive scanning would be in line with the prohibition on general monitoring in EU platform law.[72] South Korea has a more direct obligation pending at its National Assembly, reacting to a surge in fake medical ads, which falls short of requiring scanning. This would require comms platforms to provide uploaders with labeling tools, and to notify them of their obligations under South Korean law to label their content and not to tamper with watermarks and labels required by law earlier in the AI pipeline.[73]

In California, a 2024 law did require platforms to detect, label as manipulated and inauthentic, and in some cases block, deceptive election-related media, explicitly motivated by the rise of generative artificial intelligence.[74] It was struck down in 2025 with a permanent injunction after being challenged by X on the basis of breaching the U.S. intermediary liability shield, Section 230.[75]

China and India have gone the furthest in requiring platforms to respond to labels. India obliges large comms platforms designated by the government ("significant social intermediaries") to both allow users to declare whether content is synthetic, and to verify this with appropriate technical measures — perhaps by scanning for watermarks, running a detector or both — and in any case to ensure content is correctly labeled.[76] China requires comms platforms to label content as being, or possibly being, synthetic: if watermarks are detected;

71: 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 35.

72: DSA, art 8; see generally Martin Husovec, *Principles of the Digital Services Act* (Oxford University Press 2024) ch 5.

73: 정보통신망 이용촉진 및 정보보호 등에 관한 법률 일부개정법률안 [Bill to Partially Amend the Act on Promotion of Information and Communications Network Utilization and Information Protection], Bill No. 2216186, 22d Nat'l Assembly (S. Kor. Jan. 20, 2026) (introduced by Rep. Cho In-cheol [조인철] et al.).

74: Defending Democracy from Deepfake Deception Act of 2024, ch. 261, 2024 Cal. Stat. (codified at Cal. Elec. Code §§20510–20520 and Cal. Civ. Proc. Code §35).

75: *Kohls v. Bonta*, No. 2:24-cv-02527-JAM-CKD, ECF No. 98 (E.D. Cal. Aug. 20, 2025) (order granting summary judgment on AB 2655).

76: The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (India) rule 4(1A) (as amended by the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026).

if media forensics or AI detectors are triggered, implying the scanning of all uploaded content; or if the uploader or other users flag it as such, with different labels depending on the certainty.[77]

One question that these laws broadly leave unanswered is the availability and the documentation of the detectors. Only California's watermarking law explicitly requires an open detector to be made available.[78] As mentioned in Chapter 2, the watermarking systems typically rely on a hidden key that only the company issuing the watermark knows, and which is required in order to detect the mark. These kinds of hidden watermarking schemes are not new. They are reminiscent of the multi-decade, under-the-radar effort by industry and governments to add secret identifying markers to all color printing. Many manufacturers began subtly watermarking every document their printers produced with identifiers — often faint yellow dots, barely visible except under ultraviolet light — that could reveal the specific printer used and the time and date of printing. These tools have been connected to the U.S. Secret Service, and accused of being used to uncover government whistleblowers.[79] NGOs have attempted to document and reverse engineer these watermarks, as the specifications or schemes are not public. A German data protection regulator, on attempting to learn more about the legal basis for this practice during its own procurement, was stonewalled repeatedly.[80] In the AI context, the question of public disclosure around whether watermarks are being placed or used, and of who has the ability to understand what they contain, is an important if underappreciated aspect of governance in this space going forward.

CONTENT CREDENTIALS

Watermarking laws and obligations put forth by policymakers generally focus on marking out certain media as synthetic. Yet comms platforms have been closely involved in a multi-stakeholder governance project with a related, yet nonetheless notably different objective. Given the difficulties in AI content detection we've just discussed, and the likely impossibility of forcing all users everywhere to use models that have incorporated watermarking schemes, what if instead of marking the AI-generated potentially 'fake' or untrustworthy content, one could reliably mark everything else as 'real' content? This tactic isn't unheard of in the AI stack: think back to our discussion of how Cloudflare, Apple and others have developed network protocols in an effort to identify and brand the real humans, rather than doing the very difficult work of accurately identifying the fake bots (Chapter 4).

Enter the 'content credential.' In 2019, the BBC, CBC, New York Times and Microsoft

77: Measures for Labeling of AI-Generated Synthetic Content (人工智能生成合成内容标识办法), State Info. Office Tongzi [2025] No. 2 (promulgated by the Cyberspace Administration of China and others, 14 March 2025, effective 1 September 2025) (China), art 6 .

78: Cal. Bus. & Prof. Code §§ 22757–22757.6 (West 2026) (California AI Transparency Act); the European Commission however argues that the condition in the AI Act that a watermark is 'detectable' implies a requirement for a publicly accessible detector, see European Commission, Guidelines on the Implementation of the Transparency Obligations for Certain AI Systems under Article 50 of Regulation (EU) 2024/1689 (the 'AI Act', C(2026) 5054 final) (2026) para 75.

79: Derek Hawkins, "How Tech Sleuths Cracked the Mysterious Code That Turns Your Printer into a Spying Tool" (*The Washington Post*, 9 June 2017) https://www.washingtonpost.com/news/morning-mix/wp/2017/06/09/how-tech-sleuths-cracked-the-mysterious-code-that-turns-your-printer-into-a-spying-tool/ accessed 9 March 2026.

80: Unabhängiges Landeszentrum für Datenschutz, "Vorsicht: Yellow Dots! Versteckte Informationen in Farbkopien" (23 May 2019) https://www.datenschutzzentrum.de/uploads/it/2019_ULD_Report-Yellow-Dots.pdf accessed 8 March 2026.

formed 'Project Origin,' aiming to create a technical pipeline through which news organizations could authenticate shared stories and media, and indicate that certain content was indeed real.[81] That same year, Adobe announced the 'Content Authenticity Initiative,' with similar aims for users of editing software.[82] Adobe was concerned that in a scramble for authenticity, the legitimate photoshopping that the firm relies on for its survival would be stigmatized. "[I]ntent matters," the initiative's director implored. "Editing photographs is good."[83]

In 2021, these two efforts merged under the rubric of an industry consortium known as the Coalition for Content Provenance and Authenticity, or C2PA, promising a "standardized provenance solution with the goal of combating misleading content."[84] While some AI firms are members, the coalition also features camera manufacturers, news and media organizations and chipmakers.

At each stage of an image's life-cycle, C2PA proposed that different hardware and software providers create 'attestations' as to what they did to the image, and then cryptographically sign that information and include it in the metadata. A hypothetical chain (Figure 6.2) would begin with an individual taking a photograph with their C2PA-enabled camera, which would cryptographically sign the photo and its timestamp, followed by a separate attestation as to any edits carried out in an editing suite like Photoshop, followed by yet another attestation to any filters applied on Instagram or other social networks where the image would then be posted. C2PA has been seeking codification as an officially recognized international standard.[85] All images worthy of your trust should, C2PA would argue, be accompanied by this metadata.

C2PA is a metadata standard, primarily concerning what is in the metadata. Affixing the metadata to your file is a different question. Some file formats, like .jpeg or .png files, have slots suitable for such metadata built in, while others, such as .gif files, would need a separate, accompanying .c2pa file. There are some suggestions that a short identifier could be embedded as a watermark, linking to an external database of .c2pa files or information, or a distributed database could feature hashes of images as they are made, which could be queried for metadata. But both these strategies rely on Internet access at the point of attestation, and create central points of trust which could be compromised, for example, by intelligence or security agencies.[86]

Trust and security are serious problems with C2PA. C2PA is designed to be tamper-evident.

81: Jatin Aythora and others, "Multi-Stakeholder Media Provenance Management to Counter Synthetic Media Risks in News Publishing" (International Broadcasting Convention (IBC), 1 September 2020) https://iptc.org/std/MediaProvenance/Documents/Origin%20-%20IBC%20Paper.pdf accessed 9 March 2026.

82: Dana Rao, "The Content Authenticity Initiative Summit: Collaborating to Drive Trust and Transparency Online" (*Adobe Blog*, 10 February 2020) web.archive.org/web/20200701024413/https://theblog.adobe.com/the-content-authenticity-initiative-summit-collaborating-to-drive-trust-and-transparency-online/.

83: Andy Parsons, "Welcome" (*Content Authenticity Initiative*, 20 January 2020) https://contentauthenticity.org/blog/welcome-andy-parsons accessed 17 November 2025.

84: Coalition for Content Provenance and Authenticity, "C2PA Founding Press Release - Technology and Media Entities Join Forces to Create Standards Group Aimed at Building Trust in Online Content" (*C2PA*, 22 February 2021) https://web.archive.org/web/20220419011331/https://c2pa.org/post/c2pa_initial_pr/ accessed 9 March 2026.

85: International Organization for Standardization, *Authenticity of Information — Content Credentials* (Committee Draft ISO/CD 22144, ISO 2024) https://www.iso.org/standard/90726.html accessed 5 March 2026.

86: Tu Bui, Shruti Agarwal, and John Collomosse, "TrustMark: Robust Watermarking and Watermark Removal for Arbitrary Resolution Images" in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2025) https://collomosse.com/pubs/Bui-ICCV-2025.pdf; John Collomosse and Andy Parsons, "To Authenticity, and Beyond! Building Safe and Fair Generative AI Upon the Three Pillars of Provenance" (2024) 44(3) IEEE Computer Graphics and Applications 82 DOI: 10.1109/MCG.2024.3380168 accessed 8 October 2024.

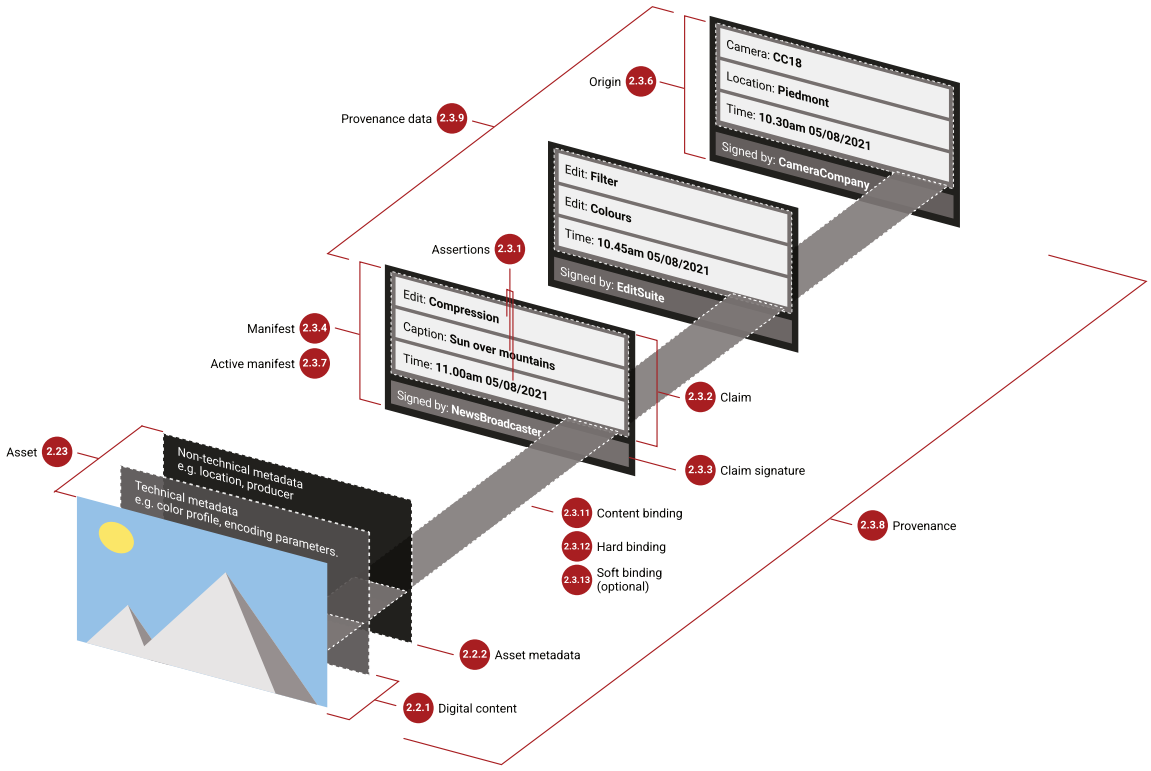

Figure 6.2: Stylized depiction of the C2PA Architecture. Image by the Coalition for Content Provenance and Authenticity, reprinted under a CC BY 4.0 license.

Someone examining the C2PA metadata should be able to see if the underlying file has changed compared to the one the metadata relates to. Yet what is to stop someone hacking a version of Photoshop to sign an image saying only minor edits have been carried out from an authentic original, when in practice, the image is barely recognizable? This is the classic 'oracle problem,' perhaps more fittingly called 'garbage in, crypto out.' These terms refer to the situation where incorrect or even malicious information is given an imprimatur of trust by a digital signature, blockchain, non-fungible token or some other improbable cryptographic trend.

To limit the risk of forgery through hacking, C2PA intends to require entities that want to make attestations to show their products meet a certain level of security. We already discussed extensively in Chapters 2 and 3 how difficult it is to control what happens on an open-ended device, especially against determined individuals. C2PA is currently standardizing two levels of security for those they authorize to issue certificates: a basic software-only level 1, and a level 2 which requires use of hardware-based attestation.[87]

Hardware attestation is our controversial friend from Chapters 3 and 4, where we explored it particularly in the context of the Apple Secure Enclave. Software- and hardware-based

87: C2PA Technical Working Group Conformance Task Force, "C2PA Generator Product Security Requirements (Version 0.1)" (2 June 2025) https://github.com/c2pa-org/conformance-public/.

attestation are radically different. Software-based security effectively means that anyone with control of the operating system can locate the key that is being used to sign the metadata and forge it themselves. Hardware-based security hides the key deep inside the device, in a special enclave that cannot be retrieved, and only gives it to an application that proves that its software has not been tampered with. A hacked version of Photoshop would, if properly designed, not qualify. Hardware-based security is controversial, as described in Chapter 3, because it is fundamentally about protecting the interests of the device against its user.

What difference does the distinction between Level 1 and Level 2 security actually make in practice? C2PA argues that the assessment of trustworthiness should be up to the beholder.[88] In practice, this means it will be largely up to the comms platform how it interprets and presents this information to users. The initial conception of C2PA was primarily user-centric: provide C2PA metadata to users, allowing them to then make a judgment about the content they're seeing. But not all devices will be capable of producing a Level 2 certificate — software running on older or no-longer-maintained devices; open-source operating systems that do not make use of the enclaves such as Linux and 'de-Googlefied' variants of Android; cameras or microphones without advanced chips. Many of the images and media that will come from controversial situations will come from such devices, such as from struggling individuals using whatever equipment they can in warzones, or from locked-down phones from individuals wary of the code they run and state interception. Level 2 certificates, just like the network authenticity tokens and the governance-by-App-Store discussed earlier in this book, give massive power to operating systems and infrastructure providers. They further give power to the C2PA and other certificate authorities, who decide who is able to sign 'authentic' images, and under what conditions.

Sometimes, individuals will own older devices that are no longer being updated. If these are successfully hacked to apply C2PA signatures to arbitrary content, either the manufacturer or C2PA would revoke the validity of the affected certificates, as Nikon did when their C2PA system was compromised in late 2025.[89] Where the manufacturer no longer updates their products, because they are too old, or because the firm no longer exists, individuals will own a device that will register as untrusted against C2PA tests. Older devices, whether no-longer-updated ones or those that only support lesser forms of hardware and software security, risk ending up as little more than e-waste.

Contrary to the hopes of the C2PA leadership, there is little evidence that individuals will know much about what to do with these trust levels and labels. Even the relatively simple case of secure HTTP has been incredibly hard for users to navigate and understand. Browsers often indicate whether a connection is made over encrypted HTTP, known as HTTPS, the protocol popularized after the Snowden leaks in order to protect browsing against network eavesdroppers, and now typically present a warning screen when an individual tries to access an unencrypted website, or one with an untrusted certificate. Like C2PA, this is foundation-

88: See statement by Leonard Rosenthol, Chair of the C2PA Technical Working Group and Senior Principal Scientist at Adobe Kate Kaye and Pam Dixon, "Privacy, Identity and Trust in C2PA: A Technical Review and Analysis of the C2PA Digital Media Provenance Framework" (*World Privacy Forum*, September 2025) https://worldprivacyforum.org/media/documents/c2pa_report.pdf accessed 9 March 2026, at 27.

89: Sebastian Oakley, "Nikon Revokes All C2PA Image Authenticity Certificates after Major Vulnerability Exposed" (*Digital Camera World*, 24 September 2025) https://www.digitalcameraworld.com/photography/photojournalism/nikon-revokes-all-c2pa-image-authenticity-certificates-after-major-vulnerability-exposed accessed 9 March 2026.

ally a question of who signs the certificate, how trustworthy they are and whether everything matches and looks to be in order. Yet all major browsers have now entirely removed or de-colored the 'padlock' symbol, as studies showed that users hugely misunderstand both its presence and absence, including often believing it represented a website they could, in general, trust.[90] Indeed, users are not wholly irrational here: just like with the C2PA scheme, there are many valid reasons that a website's security may appear poor, even when it actually is not, habituating people to just ignore the subtle warning signs and click through.[91]

C2PA's impact is dependent on what, if anything, comms platforms do with the attached information. If they only labeled it, this would require users to build a mental model of a very complex, chained system of provenance in a context where they would rarely see non-C2PA-authorized content. This matter would become even more confusing as individuals would not only have to reason about chains of edits, but also whether they should trust anything with one or more of the provenance steps only at Level 1 security. Ultimately, there is very little in C2PA alone to stop an individual from taking a high-resolution snapshot of a screen with a trusted camera and having it certified that way. Would the Frenchwoman have been convinced by a warning telling her that Brad Pitt was only talking to her through software-level C2PA provenance? Would she have even been using a social media platform so polluted with confusing labels?

Instead of labeling, comms platforms could instead choose to *only* permit media which has an appropriate C2PA manifest, either by blocking or massively downranking everything else. Even end-to-end encrypted messaging platforms could feasibly, on device, check whether the images, audio or video that users were sending to each other had some kind of manifest attached. A platform doing this would have to address many questions. How would it treat pre-C2PA content and memes? How would it address screenshots? How would it address links to a photo on the Web? Perhaps more than anything, such a universal, content-based infrastructure designed to block content could quite easily be co-opted for censorship.

There's also a political challenge: if you did want comms platforms to block all non-C2PA content, how would you force them to do this? Attempts to legislate would likely run into free expression concerns. The platform rules described above that do apply marking obligations to platforms do so to content suspected of being synthetic — not all content. A substantial industry and possibly also public backlash could be expected. As we have discussed, firms like OpenAI have little desire to support watermarking of content such as text, worried that doing so will negate the core value-add of their product — people pretending not to be using AI when interacting with the world.[92]

90: Emanuel von Zezschwitz, Serena Chen, and Emily Stark, ""It Builds Trust with the Customers" - Exploring User Perceptions of the Padlock Icon in Browser UI" (2022 IEEE Security and Privacy Workshops (SPW), May 2022) DOI: 10.1109/SPW54247.2022.9833869; David Adrian and others, "An Update on the Lock Icon" (2 May 2023) https://blog.chromium.org/2023/05/an-update-on-lock-icon.html accessed 9 March 2026; Jen Simmons and others, "WebKit Features in Safari 18.4" (31 March 2025) https://webkit.org/blog/16574/webkit-features-in-safari-18-4/ accessed 9 March 2026.

91: Devdatta Akhawe and Adrienne Porter Felt, "Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness" in *22nd USENIX Security Symposium (USENIX Security 13)* (2013) https://www.usenix.org/conference/usenixsecurity13/technical-sessions/presentation/akhawe accessed 9 March 2026; Mustafa Emre Acer and others, "Where the Wild Warnings Are: Root Causes of Chrome HTTPS Certificate Errors" in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (CCS '17, Association for Computing Machinery 30 October 2017) DOI: 10.1145/3133956.3134007.

92: See discussion in Chapter 2.

This speaks to C2PA, unfortunately, being the wrong tool for the job of stopping misuse. It may indeed add some friction and some opportunities for governance by certain platforms. Yet C2PA was established by media organizations, for media organizations, and it shows. It is certainly useful for entities looking to build a tight, cryptographically assured chain of created materials, such as press photos, which can more easily have a clear provenance chain, filled with the most secure hardware and only edited on the latest, most secure software. When someone wants assurance that something is definitely true, the gold standard provides a lot of that evidence. But when it comes to separating the authentic from the inauthentic in the messy primordial soup of the Internet, content credentials are patchy and insecure, are difficult for users to interpret, and consolidate digital power around a few hardware actors. Any ambition for C2PA to solve more than a narrow range of problems seems, unfortunately, to be a totalizing dream.

⋆

What does this all mean? There is no content-oriented detection solution that ticks all of the necessary boxes. Detection fails. Watermarking is simply avoided by malicious actors. Content credentials are hopelessly totalizing. For this reason, comms platforms have been experimenting with alternative approaches that don't require them to look at content at all.

NON-CONTENT-ORIENTED DISRUPTION

Sometimes comms platforms act assertively to disrupt the malicious behavior of their users, including by working with other actors in or outside of the AI stack. They can do this without necessarily needing to analyze content shared by users. Instead, they can analyze, at a higher level, the activity of those accounts, and the metadata associated with them.

MANUAL INVESTIGATION AND BEHAVIORAL MODERATION

The largest user-generated content platforms have, in the past decade, hired extensively from law enforcement and intelligence circles to build small, specialized teams of proactive investigators who work full-time, with extraordinary visibility across the platform, to sniff out the most harmful forms of misuse. The roots of this go back more than a decade.

In 2013, a California man logged into Facebook and used a pseudonymous account to contact a teenage girl in Indiana. Claiming that he was a friend-of-a-friend, and that he had obtained salacious photos that she had previously shared with an ex, he began a 16-month-long campaign of harassment and intimidation. After extorting multiple sexual images and videos from the teen, her efforts to eventually bring an end to their interactions would lead to bomb threats, the temporary closure of her school and various businesses in her community and, eventually, the involvement of U.S. state-level and federal law enforcement.[93]

93: Attorney's Office for the Southern District of Indiana (United States), "Buster Hernandez, Aka 'Brian Kil' Charged in Nation-Wide Cyber Sextortion and Threat Case" (*Department of Justice, United States of America*, 30 August 2017) https://www.justice.gov/usao-sdin/pr/buster-hernandez-aka-brian-kil-and-purge-maine-charged-nation-wide-cyber-sextortion-and accessed 17 November 2025.

Per court documents, 'Brian Kil' operated "hundreds of Facebook accounts" that he had set up with a wide range of pseudonyms, using them to successfully repeat similar 'sextortion' scams on other young women around the United States.[94] He was a relatively savvy operator, accessing Facebook only through the Tor network, using the Tails secure operating system designed to leave no trace on a device once shut down and verifying his profile with various 'burner' email services, making it particularly difficult for the company to detect his activity and to intervene in ways that would limit his ability to use the platform. Nevertheless, after law enforcement brought the Indiana case to Facebook, it kicked off an unprecedented two-year operation inside the firm to not only crack down on this extremely motivated and malicious bad actor, but also to help law enforcement eventually de-anonymize him and bring criminal charges.

Meta, then Facebook, expended significant time and resources in its effort to counter the activity of 'Brian Kil,' perhaps the most irritating and problematic of its more than a billion daily active users. According to some remarkable reporting published by *Motherboard*, significant organizational and technological interventions were taken by Facebook as they sought to investigate their law enforcement tip, including the appointment of a dedicated full-time staffer 'red teaming' the Facebook platform just to look for fake accounts *operated by this one individual*.[95] The company is also reported to have developed new automated systems for analyzing user metadata and behavior patterns to try to predict whether or not a newly created account was in fact operated by Brian Kil or engaging in similar behavior. The episode ended spectacularly when Facebook worked with a cybersecurity firm to develop a zero-day exploit for the Tails operating system, which they secretly delivered to the abuser via Facebook to obtain his real IP address, eventually allowing the FBI to make the arrest.[96]

Since Meta began searching the entirety of its platforms for potential 'Brian Kils' more than a decade ago, the company, as well as other large platforms, has developed various techniques and systems for identifying particularly harmful actors. In contrast to the ordinary content moderation paradigm, this is specialized work done by expert investigators, working manually (albeit with various data and analysis tools at their disposal) to look into fraudsters, traffickers and state-sponsored influence operations. Firms do this because illegal activity happens on less-visible surfaces of a platform, where they cannot count on flags from other users to draw their attention to potentially malicious accounts. Content might also be encrypted for security and privacy reasons, such as in messaging, and is thus structurally unavailable to be analyzed and flagged by automated systems.[97] Assessing 'signals' of potential malicious use — metadata such as the emails used to sign up for accounts, user IP addresses, the type of activity the user engages in, their social graph — can help flag accounts for manual investigation.[98]

These kinds of behavioral, non-content-oriented moderation techniques have not been ex-

94: *United States v. Hernandez*, No. 1:17-cr-00183 (S.D. Ind. filed Sept. 7, 2017).

95: Lorenzo Franceschi-Bicchierai, "Facebook Helped the FBI Hack a Child Predator" (*Vice Motherboard*, 10 June 2020) https://www.vice.com/en/article/facebook-helped-fbi-hack-child-predator-buster-hernandez/ accessed 17 November 2025.

96: ibid.

97: Riana Pfefferkorn, "Content-Oblivious Trust and Safety Techniques: Results from a Survey of Online Service Providers" (2022) 1(2 2) Journal of Online Trust and Safety DOI: 10.54501/jots.v1i2.14 accessed 22 March 2024.

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

tensively researched in the academic literature on platform governance. A notable exception is found in the work seeking to unearth 'coordinated inauthentic behavior' and spot state-sponsored disinformation operations.[99] Investigators inside firms, using open-source and proprietary tools for network analysis, can access far more information about accounts and about content than is available to ordinary moderators tasked with making a rapid judgment about the acceptability of an image or video.[100] This has included IP addresses, browser fingerprints, signup emails and more, but can also involve going 'off-platform' to enrich that data with information gleaned from other databases, sources or the open Web.

In a 2023 paper, Meta researchers described their efforts at what they called 'strategic network disruptions,' which they had scaled from their early efforts to battle networks of government-linked propaganda accounts to also use against "hate networks, cyber espionage and mass reporting."[101] This strategy could be understood as an extension and expansion of early efforts by firms to combat the digital activity of ISIS and other related groups, and became widely adopted by well-resourced platforms before the second Trump presidency, as they invested resources in not just reactive moderation but also proactive investigation into problematic actor groups.[102]

That paradigm has been taken on by some AI firms. After hiring staff from the research firm Graphika, which became known for the combination of network analysis and open-source investigation techniques that it deployed in the effort to unearth evidence of Russian interference in the 2016 U.S. election, OpenAI began issuing periodic 'threat intelligence reports,' in which the company briefly examines various coordinated malicious campaigns that used DALL·E or ChatGPT. The fall 2025 report discusses, for example, a

> *likely Cambodia-origin scam operation we disrupted [which] used our models to generate detailed biographies for fake investment experts and fictitious employees of fake trading firms. The scammers then asked ChatGPT to write social media messages in those characters' voices. Often, the threat actors input to our model messages they appear to have received from their targets, and asked the model to continue the conversation as the fake persona.*[103]

Many of these disclosed incidents appear to have been identified by OpenAI after receiving information from 'trusted sources,' allowing it to then pore through its logs for deeper analysis; others may have been detected after comparing user logs to the kinds of threat 'signals' that are widely shared within the cybersecurity community. Anthropic has also published detailed threat intelligence reports, for example detailing various successful efforts by Claude

99: Jen Weedon, William Nuland, and Alex Stamos, "Information Operations and Facebook" [2017] Facebook Security White Paper https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf.

100: Camille François and evelyn douek, "The Accidental Origins, Underappreciated Limits, and Enduring Promises of Platform Transparency Reporting about Information Operations" (2021) 1(1) Journal of Online Trust and Safety DOI: 10.54501/jots.v1i1.17.

101: Guy Rosen, "Raising Online Defenses Through Transparency and Collaboration" (*Meta Newsroom*, 29 August 2023) https://about.fb.com/news/2023/08/raising-online-defenses/ accessed 18 February 2026, n.p.

102: Tom G Thorley and Erin Saltman, "GIFCT Tech Trials: Combining Behavioural Signals to Surface Terrorist and Violent Extremist Content Online" (2023) 46 Studies in Conflict & Terrorism 1 DOI: 10.1080/1057610X.2023.2222901 accessed 24 January 2025.

103: Ben Nimmo and others, "Disrupting Malicious Uses of AI: An Update" (*OpenAI*, October 2025) https://perma.cc/L3XU-FPLE p. 20.

and Claude Code users to develop and deploy malware, providing some brief public detail about the filtering and investigation techniques the company had begun using to try to prevent this activity in the future.[104]

The challenge for most comms platforms is that they do not operate AI models available via API or cloud-hosted apps, and so do not also have this kind of visibility. Google can, for example, conduct detailed investigations into the use of its Gemini models, using internal logs and flagging mechanisms to manually review certain suspicious behavior[105] — and then provide hooks that other Alphabet subsidiaries and products, such as Google Search, Gmail or YouTube, can integrate into their manual investigation or automated content analysis efforts. But competitors like TikTok (which isn't also an AI developer) or Meta (which has made most of its models available in a downloadable version) don't have that same luxury. These firms — as well as firms wishing to have a more comprehensive view of potential misuse efforts, given that not all actors engaging in misuse on a certain communication platform will coincidentally also be using AI models developed by that same company — will need to coordinate to share information.

SIGNALS SHARING AND COORDINATION

User-generated content platforms have for more than a decade built socio-technical infrastructures and policy networks to share information with each other. This has been partially to help platforms make their moderation interventions more durable: open platforms like Facebook, which allow anyone to sign up with an email, have historically had little they could do to prevent actors on the spectrum of malicious use from simply coming back once their Facebook accounts were removed — that is, unless they're willing to hack and de-anonymize their users, and then sue them or refer them to law enforcement (as in the case of Brian Kil). Coordination might help other platforms also remove their services from those actors; it can also help them act more quickly.

One of the best-known examples is the Global Internet Forum to Counter Terrorism, one of the central institutional means through which the technology industry responded to the Christchurch Shooting in New Zealand in 2019.[106] It is an extensive policy network featuring some government and civil society involvement alongside many industry members with a wide range of business models (and participation from a number of actors in the AI stack discussed in this book, as well as many comms platforms). Alongside policy working groups and funding streams for research, its most salient creation is a shared hash-database, to which GIFCT member companies can upload content they have identified as breaking their terrorism policies, and which other online platforms can then automatically scan against at the point of upload.[107]

104: Alex Moix, Ken Lebedev, and Jacob Klein, "Threat Intelligence Report" (*Anthropic*, August 2025) https://perma.cc/PZ6P-FSPC.

105: Google Threat Intelligence Group, "Distillation, Experimentation, and (Continued) Integration of AI for Adversarial Use" (*Google Cloud Blog*, 12 February 2026) https://cloud.google.com/blog/topics/threat-intelligence/distillation-experimentation-integration-ai-adversarial-use.

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

107: Marguerite Borelli, "Social Media Corporations as Actors of Counter-Terrorism" (2021) 25(11) New Media & Society

The seeds of the GIFCT approach were planted with the creation of ThreatExchange, a technical infrastructure built by Facebook in 2014:

> *ThreatExchange is a simple-to-use platform that supports signal sharing among predefined groups of members in a secure, privacy-compliant, and automated way. Today, ThreatExchange (aka TX or TE) is used by multiple companies to share signals on a variety of topics intended to prevent real world harm. Some examples of how TX is currently used include sharing malware, phishing scams, and terrorism signals with the goal of helping all participating organizations tackle these problems based on their terms of service.*
>
> *Signal sharing is a tactic to prevent harm on the Internet where platforms work together to combat global threats like malware, terrorism, and other harmful content. Platforms help each other by sharing signals from content that they found and labeled on their platform. For example, Platform A might find a video of terrorism on their platform. By sharing the hash of that video (a type of signal) with Platform B, Platform B can find and review that video, which they might have otherwise missed. By sharing signals, the platforms can compound their individual trust & safety efforts and prevent more harm faster.*[108]

Today, ThreatExchange has become a platform of its own for trust and safety workers at various companies. There is a Web-based user interface where workers can access signals pertaining to removed content or accounts, and decide whether or not to keep those signals internal or share them with other companies or groups of companies, as well as a tagging infrastructure allowing them to organize these indicators by country, region and type (e.g. scams > subtype > romance). These are all available via API to be processed in an automated way and integrated into a platform's T&S workflow as they wish. Detailed information about accounts can thus be collected and uploaded to ThreatExchange in order to try to fingerprint malicious users: Google Chrome extension IDs, browsers' user agent strings, many types of domain name registration information and more.[109]

Meta open-sourced the ThreatExchange code and it now hosts the technical back-end for the Lantern consortium, a "global alliance dedicated to protecting children online," created by the Tech Coalition, an industry organization founded in 2021.[110] In 2024, 26 companies were part of this consortium, a number that has swelled to close to 50 in 2025, including virtually all major players in the North American and European market: Google, Meta, Microsoft, Apple, Amazon, TikTok, X; AI developers like OpenAI and Anthropic; niche, youth-forward services like Discord, Roblox and Snap; dating apps like Grindr and Bumble; various other platforms including Spotify, Zoom and Dropbox; infrastructure intermediaries like Cloudflare and GoDaddy; financial intermediaries like PayPal and Patreon.[111]

2877 DOI: 10.1177/14614448211035121 ; Robert Gorwa, Reuben Binns, and Christian Katzenbach, "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance" (2020) 7(1) Big Data & Society 2053951719897945 DOI: 10/ggsfrk .

108: Meta, "Get Started with ThreatExchange" (*Meta for Developers*) https://developers.facebook.com/docs/threat-exchange/getting-started/ accessed 9 March 2026.

109: Meta, "IndicatorType" (*Meta for Developers*) https://developers.facebook.com/docs/threat-exchange/reference/apis/indicator-type/v25.0/ accessed 9 March 2026.

110: Tech Coalition, "Annual Report" (2024) https://technologycoalition.org/wp-content/uploads/Tech-Coalition-Annual-Report-2024.pdf 4.

111: ibid.

The Lantern consortium helps these different actors coordinate to disrupt CSAM dissemination and creation across multiple layers of the stack. Sometimes, this action is bilateral, involving a small number of intermediaries:

> *Platform A identified a pattern: offenders repeatedly created new accounts, posted CSAM, and directed other users to contact them via specific usernames for additional material.*
>
> *Platform A shared these confirmed usernames with Platform B, which conducted investigations and then took enforcement action on 73.5% of them.*[112]

Other times, the signals go deeper and are shared more widely. For instance, the 2024 Lantern transparency report illustrates a small and yet growing number of 'GenAI prompts and keywords' that have been shared and uploaded to Lantern.[113] Imagine a scenario where a platform that is frequently used for a lot of bad-actor coordination, such as Discord, unearths some particularly malicious forums, but rather than simply removing the forum and its users, the platform shares metadata about the accounts engaged in illegal activity to support action and investigation elsewhere in the stack. When uploaded to a system like Lantern, this information might impact how investigations and governance mechanisms work at other layers we've discussed in the book, tipping off cloud or network intermediaries about certain actors and activities. Lantern metadata — such as particular jailbreaking techniques being shared via comms platforms and unearthed by investigators or flagged by platform users — could even feed back to model developers to update guardrails.

Signals sharing is more likely to be effective for the AI misuse context than classic content-oriented hash-sharing like GIFCT. AI models by definition produce new images, and have stochastic behavior, meaning that even identical prompts may produce slightly different media, making perceptual hashing an ineffective tool for combating their spread. Investigation is, however, time-consuming and usually can only enter the picture after something bad has happened. It may perhaps prevent future victims from falling prey to the same scam from the same accounts, or at least add friction and force abusers to reorganize and rethink their attacks. Through robust signal-sharing processes, attackers might not be able to easily move to neighboring platforms (or at least need to go to the trouble of using different handles, phone numbers, sign-in emails and IP addresses). Where phone numbers are required to create accounts (e.g. in the messaging context), this could create friction that might push less motivated actors to go further afield to tools that could be less effective or less commonly used for their scams. OK, you've met your new love interest online — but isn't it a bit weird that they only want to talk on Matrix, Haiket or BirdyChat?

EMERGING GOVERNANCE PATHWAYS

There are a wide range of different comms platforms with diverse business models, technical architectures and userbases, which nonetheless all play some role in hosting and disseminating malicious AI-generated content or information about how to better produce it.

112: Tech Coalition, "Lantern 2024 Transparency Report" (12 June 2025) https://technologycoalition.org/resources/lantern-2024-transparency-report/ 13.
113: ibid.

When it comes to the latter, a wide range of user-generated content services would do well to more intensively examine the ways that explicitly malicious tutorials are being spread on their services, actively red-teaming various surfaces they operate to remove the most obvious CSAM, fraud and deepfake guides. Video hosting and file-sharing sites could also consider completely banning dual-use tutorials that are not explicitly malicious — e.g. by removing some forms of instructional content that are closely affiliated with malicious use, such as faceswapping tutorials, even if those are non-sexual in nature. A platform that took this effort very seriously might change its private rules to outlaw all forms of tutorials on AI image generation. But firms would need to do this voluntarily in the U.S. and in similar jurisdictions where the jurisprudence on the lawfulness of instructional content is unclear; it's difficult to see how policymakers could force platforms to do this in a formal way without being subject to legal challenges.

A slightly different problem faces the question of disrupting activity in certain comms platforms — online forums, channel-based messaging spaces — where different forms of misuse are collectively organized. Many of these spaces are either encrypted and difficult to analyze, or operated by intentionally non-compliant actors that predicate the popularity of their platform on their decision to under-moderate and under-enforce both local laws and platform rules. Think here of Telegram. A Telegram-style platform of this type hosts countless public, semi-public or private channels devoted to the provision of illegal services, including malicious generative AI content such as deepfakes, services to commission them and models to create them. They could certainly do much more proactive investigation of certain groups and users.[114] The core challenge here is rather one of business models and platform incentives; it is a classic problem of forcing change onto a recalcitrant corporate actor, a matter which might escalate over time into criminal sanctions, penalties against firm executives and other more classic, and yet nonetheless costly and time-consuming, efforts at coercing firms to improve their private governance.[115] This can be done for the largest, most visible players, but is even more difficult to do if the platforms flee the open Web or if actors disperse to other, less commonly used platforms. Despite the challenges, perhaps the best option for regulators and security officials here is still to chase the most egregious misuse communities into corners where they are more difficult to access, increasing friction and making participation more difficult for less committed and savvy actors.

Comms platforms that are primarily implicated in the AI stack as vectors of malicious AI content dissemination — public feed-based social networks, reputable direct messaging providers — could also choose to do more to combat that dissemination. However, given the poor performance and inherent unreliability of automated techniques for content analysis, the most effective options for them are likely to be expensive, expertise-intensive forms of manual investigation and red-teaming rather than easy technological fixes.

By moving away from content and looking at signals of user behavior, many platforms are

114: Nikolaos Lykousas and Constantinos Patsakis, "Just in Plain Sight: Unveiling CSAM Distribution Campaigns on the Clear Web" [2025] APWG Symposium on Electronic Crime Research (eCrime) 1 https://ieeexplore.ieee.org/abstract/document/11327714; Sayak Saha Roy and others, "DarkGram: A Large-Scale Analysis of Cybercriminal Activity Channels on Telegram" [2025] 34th USENIX Security Symposium 4839 https://www.usenix.org/conference/usenixsecurity25/presentation/roy accessed 19 February 2026.

115: Robert Gorwa, "Durov's Arrest and the Shadow Politics of Platform Regulation" (*Tech Policy Press*, 16 September 2024) https://www.techpolicy.press/durovs-arrest-and-the-shadow-politics-of-platform-regulation/ accessed 19 February 2026.

already trying to combat fraud. This is an area where their interests generally seem to align with those of the public — scams, fraud and spam are generally bad for business. Fraud is such a lucrative proposition, however, that combating it will inherently remain a reactive game of whack-a-mole between the firms and their motivated adversaries. For the largest platforms, it occurs at such an outrageous scale that careful manual investigation cannot be the only governance mode they deploy. Services like WhatsApp are already, on average, removing close to 40,000 accounts a day under their anti-fraud policies, and heavily relying on clunky forms of automated analysis to do so.[116]

In an effort to offset the dual problems of scale and limited visibility, the general trend is toward more collaboration across various comms platforms, as well as between comms platforms and other firms in the AI stack. This is a still inchoate and emerging area, and yet already there is evidence of existing policy networks and governance infrastructures being leveraged by some firms as they combat certain forms of misuse. The Tech Coalition's Lantern infrastructure is increasingly being used to share certain malicious prompts between cloud providers, as well as hashes of content and account metadata associated with that content; there are some early examples of this infrastructure being used not just in the child safety space, but also to combat illicit deepfakes and fraud as well (as discussed in Chapter 5). The intensive use of this kind of coordination, if widely adopted, could potentially push the comms platforms away from a purely reactive mode: if, say, Google Gemini's trust and safety staff see from their logs that many users are attempting to use their services for particular forms of fraud, Google could upload the IP addresses and associated emails of those users to ThreatExchange, allowing other platforms to either ban accounts associated with those same signals or push them into specialized governance tiers — requiring them to verify their accounts in some way; placing special warnings on messages sent from those accounts (for messaging apps); escalating content posted by those accounts for more rapid human review if it is then flagged by other users (for public social networks). Unless comms platforms themselves operate cloud AI services, their ability to do the most helpful, deeper analysis is limited; they will require the providers lower in the stack to take the lead in such collaborations.

On the one hand, this kind of signals sharing seems like an obvious, important step, one that could actually have some effects in reducing the easy proliferation of certain types of malicious material. And yet, it also comes with costs. Pre-existing institutions like the Global Internet Forum to Counter Terrorism have long been critiqued by civil society and academics for being opaque 'cartels,' operating with limited transparency and accountability in the name of effectively combating malicious actors.[117] The biggest concern with powerful, informal industry governance coordination institutions is that they can be subtly shaped, or even captured, by certain stakeholder preferences and governance logics.[118] Terrorism, for example, is a hyper-political concept, and in the case of GIFCT, observers have alleged that the organization has primarily been used to respond to the kinds of terrorism associated with U.S. national secu-

116: Osmond Chia, "WhatsApp Deletes over 6.8m Accounts Linked to Scams, Meta Says" (BBC News 6 August 2025) https://www.bbc.com/news/articles/ce35q2ly1w5o accessed 19 February 2026.

117: Evelyn Douek, "The Rise of Content Cartels" (*Knight First Amendment Institute*, Columbia University, 1 February 2020) DOI: 10.2139/ssrn.3572309 accessed 5 August 2020.

118: Access Now and others, "Joint Open Letter to the Global Internet Forum to Counter Terrorism" (*Association for Progressive Communications*, 25 February 2020) https://www.apc.org/en/pubs/joint-open-letter-global-internet-forum-counter-terrorism accessed 19 February 2026.

rity concerns (namely, 'Islamic' jihadi terrorism and affiliated groups) rather than, for example, far-right and white supremacist terrorism, which may be significantly underrepresented within the GIFCT's hash-database and also general policy-networking activities[119] — ironic, if unfortunately not surprising, given that GIFCT was turned into a powerful global organization by the platform industry in response to a white supremacist mosque attack. This claim has not been tested empirically, given that GIFCT has not conducted, or at the least publicly reported on, any internal or third-party audits of the database, and the content within it.

In other words, these kinds of coordination institutions are themselves dual-use, and could be weaponized against certain types of activity or certain actors if the industry or certain firms wished to do so. They come with security, privacy and centralization concerns, and as industry coordinates, a consortium like this might itself begin to become an attractive target for hackers or government actors seeking information about not just the trust and safety operations of certain firms but also the activity of certain users. Lantern's 2024 Transparency Report relatedly notes that governments are beginning to issue requests to companies for Lantern information, and that they are developing a process for dealing with such requests.[120] Because these organizations to date have arisen in an ad hoc, informal fashion, they tend to have weak institutional safeguards and little public-facing transparency, relying primarily on corporate norms to prevent them from being misused by the platforms or governments. The same is true of some of the technical scanning and detection infrastructures we have discussed here. The same technological infrastructure that is being installed on Chinese smartphones for laudable anti-fraud purposes could quite easily be repurposed for censorship and surveillance — and in the Chinese context, some might argue that it already is.

The question of how the largest comms platforms will actually respond to the ensuing pressure from societal stakeholders remains unclear. Interestingly, while these platforms may have the most limited ability to actually meaningfully disrupt malicious AI use — being, in many cases, the space of 'last mile delivery' once that content has already been created — they are likely to be the most visible targets of public outrage and ensuing policy pressure as genAI adoption continues to rise. Time will tell how they respond to this paradoxical situation, but perhaps more drastic measures will be rolled out by industry — such as requiring all material uploaded to the platforms to go through *ex ante* rather than *ex post* manual review, or requiring all material uploaded to have some sort of secure metadata associated with it. But these firms are unlikely to adopt such measures voluntarily, given that they seriously threaten the kinds of seamless experiences that these firms have historically sought to cultivate in their efforts to capture economic value from their users and keep them on platform. Smart policy interventions and the strategic deployment of political capital by canny regulators, perhaps coordinating across jurisdictions, would in this case be needed.

119: Paresh Dave, "Inside Two Years of Turmoil at Big Tech's Anti-Terrorism Group" [2024] Wired https://www.wired.com/story/gifct-x-meta-youtube-microsoft-anti-terrorism-big-tech-turmoil/ accessed 19 February 2026.

120: Tech Coalition, "Lantern 2024 Transparency Report" (n 112) 11.