

# Letters from the Stochastic Penal Colony 🌴

When Algorithms Discipline and Punish

ROBERT GRIMM, Independent Investigator, United States

This paper serves as pointed critique of algorithmic practice outside the criminal justice system. Far too many interventions including social media's content moderation are excessively punitive, often resulting in the figurative death of users through permanent account suspension. First, based on my own experiences and grounded in procedural justice, this paper starts by exploring the many ways policy and automated enforcement turn punitive on the example of OpenAI's DALL•E 2. Second, it illustrates how even best-practices policy turns punitive performance on the example of pre-Musk Twitter. Third, a comprehensive survey of non-Chinese social media demonstrates the pervasiveness of excessively punitive content moderation. It also tests the limits of their accountability, notably by projecting the likely impact of the European Union's Digital Services Act and by correlating data released by Facebook, Google, and the National Center for Missing and Exploited Children. Fourth, to illustrate the limits of algorithmic content moderation, this paper presents a successful strategy for subverting DALL•E's aggressive automated censor, which inadvertently also unleashed grotesquely racist imagery. Fifth, this paper proposes a new intellectual property regime specifically for AI. It re-combines proven elements from copyright and patent law, resulting in a framework that balances the interests of those who invest in state-of-the-art AI and everyone else. Finally, this paper concludes by pointing towards harm reduction as a mindset for, possibly maybe, making life in this digital penal colony at least somewhat bearable—because, I fear, we are stuck in it.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; • **Social and professional topics** → **Computing / technology policy**.

Additional Key Words and Phrases: AI image generator, child sexual abuse material, content moderation, DALL•E, generative model, governance, harm reduction, Kafkaesque co-factor, punitive algorithm, social media, text-to-image tool, transparency reporting

“As you see, it consists of three parts. With the passage of time certain popular names have been developed for each of these parts. The one underneath is called the Bed, the upper one is called the Inscraper, and here in the middle, this moving part is called the Harrow. [...]

As soon as the man is strapped in securely, the Bed is set in motion. It quivers with tiny, very rapid oscillations from side to side and up and down simultaneously. [...] Only with our Bed all movements are precisely calibrated, for they must be meticulously coordinated with the movements of the Harrow. But it's the Harrow which has the job of actually carrying out the sentence. [...]

The law which a condemned man has violated is inscribed on his body with the Harrow.”

---

Franz Kafka, *In the Penal Colony* [163]

## 1 INTRODUCTION

My first interaction with DALL•E 2, OpenAI's headline-making text-to-image system, in late July 2022 didn't quite go as expected. I had signed up with OpenAI several months before but had been granted access only earlier that day. So I was eager to try out the system and started with a prompt that had yielded good results with another text-to-image system. But instead of producing four images, DALL•E 2 responded with a stern warning:

“It looks like this request may not follow our content policy. Further policy violations may lead to an automatic suspension of your account.”

Whoa! I enter an entirely reasonable prompt and OpenAI immediately threatens me with account suspension? Without even telling me what I did wrong? That’s just ridiculous. And frustrating. Consulting the content policy didn’t help much either. The policy seemed broad, also vague. I eventually did narrow down the violation to two candidates out of the eleven content prohibitions. But which of the two? I couldn’t tell. It all seemed rather Kafkaesque!

It also was familiar. In October 2021, Twitter’s AI took offense to an admittedly caustic tweet I had just posted and locked down my account. A first appeal was rejected without filling in the placeholders of the email notification template and hence without any justification. A second appeal was still pending three weeks later. All along, my Twitter experience was reduced to nothing but that same tweet, nicely centered on screen. At that point, I had enough of single tweet limbo and did submit to Twitter’s penance rites: I withdrew my appeal, acknowledged that I “violated the Twitter Rules,” and deleted the offending tweet—all with one click on the red “Delete” button.

While the two interventions were substantially different, I experienced the respective enforcement processes as similarly punitive, with little consideration given to my voice, agency, and dignity. Moreover, my prompt to DALL•E was intercepted before generating images; generated images, in turn, are not publicly visible. Similarly, my tweet was taken down within a couple of seconds after posting at most. Given the limited reach of my Twitter account, which had barely 140 followers at the time, and the early hour, just before 6 am on the East coast, I very much doubt that anyone ever saw the tweet—besides me. In short, thanks to algorithmic content moderation no-one could have been harmed by prompt or tweet. But that also renders any justification for meting out punishment null and void.

When DALL•E threatened me with account suspension, I immediately recognized the punitive thrust shared with Twitter’s content moderation and became curious about the extent of such punitive algorithmic interventions. This paper is a first attempt to map just that extent. In doing so, I focus on content moderation, which arguably is *the* business value proposition of social media [201, 245] and lately of considerable public interest.

At the same time, there is no shortage of effectively punitive algorithmic interventions. Examples include credit scoring [16], debt assessment [355], exam proctoring [97], fraud prevention [173], grading [177], job and school applications [15, 121, 122, 310], personal vendettas [48], private security services [128], productivity monitoring [62, 126, 165, 273], and screening for child sexual abuse materials or CSAM [19]. Still, Amazon’s warehouses stand out for their profit-driven amorality [166, 185]: The firm’s ruthless algorithmic exploitation of its workers not only leads to high injury rates [38, 58, 276], but the resulting 150% yearly staff turnover means that Amazon will run out of people to exploit over the next few years [277].

I call the conceptual space of punitive algorithms the *stochastic penal colony*. To provide a first map of that space, this paper first explores OpenAI’s enforcement process in §2 and then Twitter’s in §3. For each firm, it leverages a close textual reading of firm policies and other communications to enumerate the many ways the firm’s enforcement process deprives targeted users of voice, agency, and dignity. It also contrasts these *Kafkaesque co-factors* with major failures of neutral and trustworthy content moderation.

§4 shifts from auto-ethnographic case studies based on procedural justice [327–329] to a cumulative perspective, performing a comprehensive comparison of popular non-Chinese social media and their governance based on their transparency reports. The results in §4.1 suggest that content moderation is indeed excessively punitive across all social media platforms. Telegram is the only exception, but only because it doesn’t perform meaningful content moderation. Furthermore, social media governance still is far from accountable, with transparency reports addressing zero to twelve criteria out of the nineteen I identified. §4.2 explores the validity of transparency data by comparing disclosures by Google, Meta, and the National Center for Missing and Exploited Children and §4.3 uses the example of industry-leading Meta for illustrating the devastating human harm that can result from hypergrowth as business strategy. Alas, social

media’s lack of transparency and accountability may just come to an end within the year. The EU’s Digital Services Act provides a future baseline by enumerating the metrics internet platforms must include in their independently audited, yearly reports [86]. For full disclosure, I worked for Meta as a software engineer from summer 2018 to summer 2019. In the second half of 2022, I also served as paid consultant to litigation against Meta. Neither employment nor litigation were related to the topics covered in this paper and I have *no* financial interest in the company (nor any of the other companies covered in this paper).

Whereas the census in §4 takes a broad view, §5 goes deep again by experimentally exploring the reason for my first prompt’s rejection as well as my successful strategy for circumventing DALL•E’s content moderation. Since I did seek to create violative content related to the topics of this paper but have little tolerance for realistic gore, I restricted myself to images that would be suitable as editorial content for a magazine version of this paper. Thanks to that restriction, images are not substantially more disturbing than a Francis Bacon painting and I include highlights in Appendix C starting on page 54.

The hands-on investigation of DALL•E surepitiously surfaced two additional limitations of OpenAI’s content policy and automated enforcement. First, DALL•E effectively discriminated against Christian religious beliefs, simultaneously censoring too little and too much, with the latter including the crucifixion. Between me running these experiments in August through October 2022 and completing this paper in March 2023, OpenAI did update its censor and now accepts the crucifixion and hence is not as discriminatory anymore. Second, OpenAI’s post-processing of prompts to diversify gender and race representation is outright dangerous [240, 304]. In particular, it appears to be the main trigger for some prompts yielding grotesquely racist imagery. Worse, once I knew what to look for, I recognized similar, though less pronounced effects in earlier generations. These images are *not* reproduced in this paper or its appendices; though they are included with the supplemental materials at <https://github.com/apparebit/penal-colony>.

In §6, I pull together results from previous sections, organize them into larger themes, and contextualize them with related work. I argue that, even though all of the surveyed platforms with exception of TikTok launched *between one and two decades ago*, their governance remains a shambles. I also discuss solution approaches and propose a novel intellectual property regime specifically tailored to AI systems that recombines proven elements from copyright and patent law to balance the interests of those who deploy such systems and everyone else. Finally, §7 concludes this paper with an appeal to the one principle that holds promise for ameliorating life in the stochastic penal colony, harm reduction. Appendices A–D reproduce relevant source materials, whereas Appendix E discusses research ethics, including my handling of the racist imagery.

Before diving into the substance of this pointed critique of algorithmic practice, §1.1 covers the conceptual, political, historical, and literary context of the stochastic penal colony and §1.2 grounds my narrative technique for identifying punitive aspects of content moderation, *Kafkaesque co-factors*, in procedural justice. Coincidentally, that also enables me to provide a more precise definition of the stochastic penal colony.

## 1.1 Conceptual, Political, Historical, and Literary Context

In *Discipline and Punish*, Michel Foucault traces the transition from punishment as a public and usually deadly spectacle to the modern prison and other disciplinary institutions [96]. He argues that this transition did not happen for humanist concerns, as the result of reform efforts. Instead, the driving force was the destabilizing impact of public executions. By being rather ostentatious displays of power, they turned the criminal into sympathetic victim. In contrast, executioner as well as sovereign became targets of popular resentment. By simultaneously rationalizing, tempering, and distributing the application of power, penal institutions avoid these downsides. They instead instill discipline into the individual

under their custody. As people internalize discipline, that self-discipline obviates the need for more direct applications of power and begets other disciplinary institutions, including schools, hospitals, and factories. That way, we all turned into seasoned practitioners of discipline.

The institutionalization of discipline does place constraints on the sovereign's exercise of power and leads to an attendant loss of centralized control. That last aspect is often lost on Foucault's readers. They are so dazzle-dazzled by Foucault's (admittedly fascinating) idealized disciplinary institution, the panopticon, they don't realize that such a circular arrangement of cells around a central monitoring station or tower simply can't scale beyond maybe 500 cells—at least in a domain where gravity, mass, stress, and strain reign supreme. Arguably, East Germany was a notable exception. But its lo-tech approach to central control also was too expensive to be sustainable [286]. For instance, in one district, 18% of the population were active informants for the Stasi, that country's vicious state security service [169].

Here too, computing technology is proving to be a game changer—for the worse: China under its current, particularly authoritarian president Xi Jinping—or more concisely Xina—is reaping the benefits of readily available hardware sensors and machine learning algorithms. By rolling out ever more intrusive yet centralized control, Xina is erasing the distinction between prison and not-prison at scale [114, 216, 303]. Ironically, some of that is driven by American innovations on predictive policing [253, 303, 307]. But the excessive intrusiveness of the Han née Borg declaring “Resistance is futile. You will be assimilated!” also makes Xina's *surveillance state* an unsuitable model for algorithmic control in western democracies.

The *carceral state* in the United States comes closer [298]. The country has the largest number of prisoners *and* the highest incarceration rate in the world. In 2022, the country accounted for 20% of imprisoned people across the world [280] but only for 4.25% of the world population [354]. With 625 per 100,000 in 2019, the us imprisoned 6.0× as many people as Canada, 9.4× as many as Germany, and 17.5× as many as Japan [353]. Additionally, roughly double as many people are under the control of the carceral state through probation or parole and may be locked up again at a moment's notice for largely arbitrary reasons [280]. Many of the imprisoned are deprived of their liberty not because they have been convicted of a crime: Half of those imprisoned in local jails (as opposed to federal prisons) are there for pretrial detention [280]. Worse, mass incarceration disproportionately impacts poor and minority populations. Notably, Black Americans made up 12% of the us adult population in 2018 but also made up 33% of all people serving sentences greater than one year in state or federal prison [110]. Incredibly, that is after a 34% reduction of their incarceration rate since 2006 and comparatively smaller reductions for other ethnicities. They also tend to be locked up, far from home, in districts that are far more white than their homes, cutting them off from family and friends [338].

As I'll discuss in more detail in Section 4, some of the more extreme and hence also white supremacist practices of the American carceral state, notably the (figurative) death penalty and three strikes rules, have direct equivalents in social media content moderation. In the opposite direction, the carceral state is an early and aggressive adopter of algorithmic enforcement technologies [12, 83, 124, 265, 356]. At the same time, algorithmic enforcement clearly isn't limited to the government. Corporations and universities are deploying punitive interventions just as well outside the criminal injustice system. In doing so, they may even innovate on punishment. As I'll show in §3, Twitter's enforcement process harkens back to punishment as a performance, but it does so while (ingeniously) avoiding the destabilizing public spectacle. In short, this paper is concerned with algorithmic interventions outside of the immediate control of the sovereign state. That also distinguishes this work from previous papers, which point towards the punitive potential but do so in the context of the surveillance and carceral states [72, 204].

I am proposing the *penal colony* as closest historical precedent and as fitting model for contemporary algorithmic practices outside the criminal injustice system. The French version of *transportation*—the practice of sending prisoners to

far off locales—is far more recent than we’d probably like to acknowledge. France began turning French Guiana into one large penal colony from 1852 onwards—after the British had already begun unwinding their own penal colonies—and closed the colony only in 1953 [5, 10, 306]. For the 70 years or so before closing, transportation was reserved for convicts sentenced under France’s own three strikes laws. It also was almost always terminal: Only 2,000 out of 70,000 prisoners returned to France during their lifetimes [340]. For that reason, prisoners referred to the penal colony as “dry guillotine” [100, 268]. Yet discipline was surprisingly inconsistent, even lax, depending on location.

Foucault had surprisingly little to say about the penal colony [266], even though transportation must be understood as a distinct intermediate stage in penal history. As such, it combines aspects from the performance of punishment and the discipline of prisons. Notably, like earlier practices, transportation is usually terminal. But unlike earlier practices, the penal colony is discretely out of sight. The penal colony also incorporates a disciplinary component, typically involving hands-on labor to create the infrastructure for more general colonization. The *stochastic penal colony* stands apart from both carceral and surveillance states, too. Its focus on inhuman punishment does remind of the American carceral state, but its technology is much the same as Xina’s surveillance state. Yet, there is no central control, or even intent. Its downsides, not surprisingly, worsen along racial lines, but it ensnares the privileged, including many White people, almost as easily.

Finally, the remoteness of the penal colony, both literally and figuratively, also turns it into an effective investigative device that renders contemporary practice strange again and hence amenable to analysis. While that renders the stochastic penal colony a largely ahistorical concept, its intellectual lineage does trace right back to the former penal colony in French Guiana: The stochastic penal colony obviously draws on Franz Kafka’s 1919 short story *In the Penal Colony* [163]. Kafka, in turn, was influenced [270] by Octave Mirbeau’s 1899 novel *The Torture Garden* [211]. While taking place in an imaginary China, its year of publication and dedication—*To Priests, Soldiers, Judges / to men who rear, lead, or govern men / I dedicate these pages of murder and blood.*—point to the Dreyfus affair as primary inspiration. Alfred Dreyfus, a Jew and French military officer, had been falsely convicted for espionage in early 1895 and again in 1899—with rampant antisemitism leading to the systematic suppression of exculpatory evidence and complete disregard of the real spy’s public confession in 1898. As a result, Mr Dreyfus spent 1895–1899 on Devil’s Island, a particularly harsh location in the French Guianan penal colony. Coincidentally, the Dreyfus affair also popularized the word “intellectual,” albeit starting out as a pejorative [77, 315].

Besides, the stochastic penal colony 🌴 provides an excellent home for a pandemonium of stochastic parrots 🦜 [24]!

## 1.2 Foundational Criteria

In the auto-ethnographic case studies in the following two sections, I focus on the procedural aspects of policy enforcement. In part, that reflects the fact that the outcomes aren’t particularly interesting. I still have both accounts and they both still work. In part, that reflects the fact that I, just like the people observed by Tyler and others, evaluate the justice of an intervention based on its process. That endorsement is in fact based on an inadvertent experimental outcome: I only learned of Tyler’s work after having performed a basic procedural justice analysis.

As already mentioned above, my analysis labels violations of a user’s voice, agency, and dignity as Kafkaesque co-factors and also considers violations of neutrality and trust while considering additional sources. Procedural justice positions voice and neutrality for evaluating the process itself, while respect and trust are positioned for evaluating the relationships between participants. If we disregard my playful labelling of “Kafkaesque co-factors” and rename “dignity” with “respect,” the similarities are obvious.

So is the primary difference: my inclusion of agency in addition to voice. Voice means being listened to and given due consideration. Agency means the freedom to make one's own decisions and then act on them, with an emphasis on the freedom to act part. In other words, agency implies initiative whereas voice does not. The omission of agency from procedural justice isn't too surprising when we consider its original context, namely sovereign law and justice. In the governmental application of justice and particularly in criminal justice, individuals' agency is at best a secondary concern and for the most part lacking. As the above summary of outcomes for the American application of injustice demonstrates, even voice is on exceedingly shaky grounds in the us.

A more subtle difference is my grouping of the five criteria. Whereas procedural justice distinguishes between process and relational criteria, I distinguish between criteria that stand on their own and hence are meaningful even when applied to just a single governance process, i.e., voice, agency, and dignity, as well as criteria that require additional context, i.e., neutrality and trust. Clearly, a negative individual experience will influence one's evaluation of neutrality and trust as well. But being confident in statements about neutrality and trust when it comes to governance procedures fundamentally requires more than just one exemplar.

Having clarified the differences in criteria, I can now also give a better definition of the stochastic penal colony: It's the universe of algorithmic interventions outside the governmental application of justice that routinely violate voice, agency, dignity, neutrality, and trust. That in turn also helps clarify the distinction from Foucault's disciplinary institutions: The latter imply some compromises when it comes to voice, agency, and dignity. But in the modern conception, they also require neutrality and trust. Routine violations of the latter two turn a disciplinary intervention into a penal one. For those same reasons, the American carceral state is not a disciplinary institution in line with Foucault's understanding but something altogether worse.

## 2 DALL•E 2 SUPERMAX

Back to OpenAI and its warning: As stated in the first sentence—"It looks like this request may not follow our content policy."—the alleged policy violation is just that, some violation of some policy. The warning's complete lack of specificity directly translates into a lack of actionable guidance—beyond obviously not submitting the prompt again. At best, that seems just lazy. Like OpenAI couldn't be bothered to write a better warning. At worst, it feels condescending: We are being treated as children—or pets. That is the first Kafkaesque co-factor.

The first sentence does link to OpenAI's content policy (which I reproduced in Appendix A.1 on page 49). Let's see whether the policy is of more help. We click the link and, a moment later, our eyes are immediately drawn to the policy's third, bold-faced sentence. It commands:

**“Do not attempt to create, upload, or share images that are not G-rated or that could cause harm.”**

That command is followed by eleven bulleted and also bold-faced categories—hate, harassment, violence, self-harm, sexual, shocking, illegal activity, deception, political, public and personal health, and spam. The eleven categories' jumble of nouns and adjectives is followed by two more commands and a request, each with its own bulleted items—disclose the role of AI, respect the rights of others, and please report any suspected violations.

All four directives have the same basic structure and visual layout, a bold-faced exhortion followed by bulleted elaboration. Typically, that would suggest that they are of similar importance. But when also considering their ordering, the number of bulleted items (eleven, two, three, and one, respectively), and the presence of bold-faced text amongst the bulleted items (eleven, none, none, and none), the first directive clearly is the most important. Judging by that same presentation, the eleven categories elaborate on that prime directive.



The eleven categories also make the content policy expansive. While it is straight-forward enough to come up with examples for harmful content in each of the eleven categories, claims that, say, political or health-related content is necessarily or even mostly harmful seem rather preposterous. The policy also is very much prescriptive: It commandeers without providing motivation or justification. Context, which is critical when evaluating marginal speech, plays no role. As a result, users cannot make informed decisions about whether content complies with the policy. Pretty much their only safe option is to stay well clear of content that might remotely touch upon the eleven categories.

The impact of the expansive and prescriptive prohibitions is twofold. First, they deprive users of agency while also limiting DALL•E's utility. Second, they create more uncertainty, especially for marginal content, when they really should be reducing it. These also are the second and third Kafkaesque co-factors.

If the warning's first sentence lacks actionable guidance and the content policy deprives users of agency while also bolstering uncertainty, the second sentence—"Further policy violations may lead to an automatic suspension of your account."—makes it all worse by raising the stakes to all or nothing—well, as far as having access to DALL•E with a particular email account, credit card, and telephone number are concerned (with OpenAI rejecting virtual numbers such as Google Voice). That qualification is an important one. It makes clear that the absolute stakes are nothing like those in Kafka's *In the Penal Colony* or also *The Trial* [164], where some unspecified transgression warrants terminal punishment. Yet the relative stakes are surprisingly similar. That makes the warning fundamentally punitive, which also is the fourth Kafkaesque co-factor.

The warning's epistemic dissonance further exacerbates its punitive character. The first sentence states that the "request *may not* follow our content policy" (emphasis mine). That is substantially different from "does not" in that it explicitly allows for misclassification. Yet the second sentence already threatens account suspension, which is both excessive and excessively punitive. Technically, that sentence and its threatened suspension are moderated by a potential "may" too. But that is too little, too late for counteracting the sentence's very excessiveness. That also is the fifth Kafkaesque co-factor.

Since my first prompt immediately triggered the above warning, we already know that use of DALL•E, an algorithm, is being monitored by another automated process. If you peruse the DALL•E subreddit (i.e., a group on Reddit), you will find that the automated monitoring process also follows through on the threat of account suspension, i.e., serves as automated censor or enforcer as well. In particular, incurring tens of warnings in as many minutes seems to be a sure way of having one's account terminated. That implies that sustained efforts at charting the boundaries of OpenAI's content policy are likely to fail with banishment as well—which further strengthens the expansive overreach of the content policy's prohibitions. That is our sixth Kafkaesque co-factor.

Issues filed in DALL•E's GitHub repository and discussion on the subreddit in August 2022 seemed to assume that policy enforcement was based on lists of words. OpenAI may have indirectly encouraged that because its safety best practices documentation at the time pointed to a blacklist. But my own experiments convincingly pointed towards a more sophisticated censor, that is, another AI. That we have reached a point where an algorithm serves as gatekeeper for another algorithm is nothing short of remarkable and only underlines how powerful stochastic algorithms have already become. Since DALL•E's gatekeeper is a language-based algorithm, likely some model derived from GPT-3 [39], we might be tempted to call it a stochastic parrot 🦜 [24]. But that doesn't do justice to its function as judge, jury, and executioner. That an algorithm has been bestowed with that awesome trinity of powers is our seventh Kafkaesque co-factor.

Independent of whether the enforcer is human or algorithmic, eventually a prohibited prompt will slip through. OpenAI appears to be well-prepared even for that eventuality: As stated in an addendum to its terms-of-use (which I

reproduced in Appendix A.2 on page 49), the firm retains ownership of all generated images. It grants exclusive usage rights, including for commercial purposes, and also promises to neither resell images nor assert copyright—“all provided that you comply with these terms and our Content Policy.” The terms continue: “If you violate our terms or Content Policy, you will lose rights to use Generations,” meaning images. That failsafe also is the eighth Kafkaesque co-factor.

In real life, OpenAI faces two complicating factors when trying to trigger the failsafe. First, us copyright law requires substantial human input for a work to be copyrightable and a prompt by itself typically won’t have sufficient substance. That doesn’t prevent OpenAI from asserting ownership. Contract law might suffice: I asked that question in Avvo’s Q&A section and the seven lawyers, who answered, cover the range from doubtful to certainly [118]. But the fact that images are not copyrightable by default does eliminate powerful instruments for limiting the spread of copyrighted content, notably the DMCA takedown notice and Copyright Claims Board [51]. The former helps with removing copyrighted content hosted by American firms and the latter makes for much faster decisions than the courts. As added bonus, neither requires a lawyer. Alas, OpenAI would have to seek a court order instead.

Second and more fundamentally, OpenAI needs to find out about the violative image and its provenance from DALL•E before it can do anything about it. According to OpenAI, the system produced over 2 million images per day late September [238] and over 4 million per day early November [237]. But even though OpenAI has access to all generations, the sheer volume also makes human quality control—beyond sampling a small number of images—exceedingly hard.

Remarkably, OpenAI seems to have anticipated this problem as well. The fourth and final directive of its content policy does not commandeer and instead starts with a polite “Please.” But then it inappropriately invites users to “report any suspected violations” and promises to “take action accordingly, up to and including terminating the violating account.” This open call for users, who themselves are subject to OpenAI’s aggressive monitoring and enforcement, to join that same effort as informants is a staple of the worst modern autocracies. It also is our ninth and final Kafkaesque co-factor.

In summary, OpenAI’s content policy and terms-of-use seek to impose exceedingly broad prohibitions on its users without any justification. Their aggressive, automated enforcement strips users of agency to make their own decisions about what content is appropriate in what context and instead treats them with punitive contempt. If that wasn’t enough, OpenAI retains ownership rights, licensing image use only, and actively recruits users to serve as informants.

The overall effect is maximally Kafkaesque. In fact, we seem to have pushed right past the limits of a hermeneutics based on Kafka’s oeuvre. After all, our AI-based enforcer is fundamentally more powerful than the machine from *In the Penal Colony*. The latter still relies on humans to do the judging and jurying, thereby allowing at least some humans to take on roles other than Condemned. But in our case, an algorithm acts as judge, jury, and executioner. It has successfully exorcised even the last vestige of humanity from a world in thrall to automated processes and quantitative everything. In short, now we all are just Condemned, waiting for the harrow of the stochastic enforcer.

## 2.1 OpenAI’s Neutrality and Trustworthiness

Before DALL•E 2’s beta opened up the system to users like me, it was available to a much smaller number of users, more like a research experiment. The system card [115, 256] for the original release in April 2022 makes clear that DALL•E 2 was, in part, trained with “publicly available sources” [212], which in all likelihood includes Internet-sourced data similar to the LAION-400M dataset [287]. While OpenAI has declined to elaborate on the exact sources for DALL•E’s training data, we know that such Internet-sourced datasets are anything but safe [28]. That makes DALL•E unsafe by design.



As shown in Appendix A.1 on page 49, DALL•E’s content policy comes right out against anything “that could cause harm.” In their FAQ entry for DALL•E’s warnings [218], OpenAI claims that “safe usage of the platform is our highest priority.” Yet DALL•E’s content policy disallows *all* political and health content. Those prohibitions are not just unusually broad, they also run directly counter the public interest in a democracy. At the same time, health, like politics, has become exceedingly partisan during the pandemic. That suggests a very specific kind of harm OpenAI seeks protection from—harm to its own reputation.

When OpenAI gave up control over generations in early November 2022, the email announcement justified that change with “improvements in our safety systems” [241]. That may be the case for generations created after those improvements were made. But when I asked customer support about generations made before the announcement, they confirmed that the new terms-of-use apply retroactively. In fact, OpenAI deleted the webpage with the terms-of-use addendum for DALL•E. But if DALL•E’s censor required improvements, then chances are that at least some older images are unsafe. Otherwise, there would have been no need for further safety improvements. Yet OpenAI pretends it can have it both ways.

What else might have put OpenAI under significant pressure to relinquish ownership in generations? What about Stability AI releasing not just source code but also model weights for Stable Diffusion in August 2022 [309]? Unlike DALL•E, that public release enables anyone with basic fluency in Python and access to recent graphics cards by Nvidia to open their own competitor to OpenAI’s system, without even paying a license fee to Stability AI. While that is pure speculation on my part, Microsoft’s \$10 billion investment in OpenAI in January 2023 on top of an earlier \$1 billion investment illustrates the stakes at play [22], which are a powerful motivation to cut corners.

In summary, OpenAI surely employs stellar public relations and legal talent. The firm also is exceedingly strategic about policies and information released to the public. Alas, the firm’s apparent need for controlling everything DALL•E is overbearing bordering on the arrogant or patrician, even though some of its positions are blatantly hypocritical. These impressions seem in line with Karen Hao’s observations in a 2020 portrait of the firm [125], which focused on the tensions between OpenAI’s founding as a research lab and its current increasingly commercial activities. However, more recent reporting about the firm’s use of labor in Latin America and Africa for labelling content is alarming [127, 252]. If confirmed, OpenAI engaged in outsourcing practices that directly harmed people and probably also violated us law. That raises grave concerns about the firm’s ethics and its ability to follow through on its ambitious charter [235].

### 3 TWEET-DA-FÉ

Early one morning in October 2021, I had just finished reading an article about some oil industry association spending millions of dollars on lobbying and advertising to derail the Biden administration’s push for climate change legislation [319]. Additionally, three of the association’s larger member companies spent millions of dollars each towards that same goal—despite also being responsible for 8.7% of all global CO<sub>2</sub> emissions since 1965 [321]. I was enraged. To vent, I composed a caustic tweet that @-mentioned the three firms and stated that I was looking forward to their CEOs facing capital punishment for genocide. I was well aware of the statement’s severity and incivility while writing it. But I told myself that that was ok, since the statement implies a formal, legal process that is still practiced in the United States. I even included that argument in both of my appeals.

I remain ambivalent about the tweet. With ExxonMobil’s internal projections from 1977 to 2003 “accurately forecasting warming that is consistent with subsequent observations” [318], with birds falling dead from the skies [67], 11 billion crabs just vanishing [233], and a third of Pakistan flooding [57] because of climate change, it is hard *not* to wish harm on responsible parties including all oil companies and their CEOs. Yet, the record-setting execution spree towards the end of

Donald Trump’s presidency [17, 171, 316] also makes clear that the content of the tweet isn’t just utterly incompatible with this paper’s basic premise, but that tweet and oil company greed share the same basic inhumanity. It appears that I acted as a *daily active shithead* that morning [295].

The tweet’s incivility certainly triggered Twitter’s AI. Within a couple of seconds after posting at most, it removed the tweet and locked my account. Compared to the content warning for DALL•E, the stated justification was far more specific:

**“Violating our rules against abuse and harassment.**

You may not engage in the targeted harassment of someone, or incite other people to do so. This includes wishing or hoping that someone experiences physical harm.”

The linked policy on abusive behavior, reproduced in Appendix B.1 on page 51, is not only specific but genuinely helpful. It is written in accessible, well-structured prose: The policy starts with a rationale, is followed by the different kinds of abusive content, and concludes with a range of possible sanctions. The mid-section on kinds of abusive content features well-delineated and reasonable prohibitions. It even reassures readers that the firm is well aware that some tweets, by themselves, may appear to violate the policy but, when considered in their original context, do not.

Thanks to the effective presentation, finding the concrete prohibition applicable to my tweet was easy: “Wishing, hoping, or calling for serious harm on a person or group of people.” After elaborating on possible context and giving examples, the policy—rather reasonably—allows that some wishes of harm may be justified, in the heat of the moment, as expressions of outrage. In such cases, Twitter still requires offending tweets to be deleted but does not impose penalties. Apparently, rapists and child abusers count as legitimate targets but oil company CEOs do not—yet.

I appealed the decision by Twitter’s AI that same morning. Or at least, I tried to: Twitter’s form for filing an appeal seems to have the same character limit as a tweet. That excludes most arguments besides a succinctly stated single reason. Alas, my justification was far from that and, not surprisingly, Twitter rejected the appeal three days later. However, the form email notifying me of the rejection wasn’t even filled in, despite containing instructions in HTML comments. Since I had located another page for launching an appeal that wasn’t marred by the original form’s character limit, I tried again with that form, this time focusing mostly on the bad form. When that second appeal went unanswered for three weeks, I gave up. I withdrew my appeal, acknowledged that I “violated the Twitter Rules,” and deleted the offending tweet—all with one click on the red “Delete” button.

Alas, residual effects from the episode remain. When I try to sign up to Twitter for Professionals, I get a notification that “something’s missing,” even though my account meets all criteria stated in Twitter’s documentation. Meanwhile, a satirical account of mine, which I opened more recently and which describes my alter ego as a “lifelong practitioner of faggotry, promoter of the gay agenda, and unrepentant socialist monarchist,” could sign up to Twitter for Professionals within days of account creation.

### 3.1 A Punishing Performance

With OpenAI’s policy enforcement being consistently Kafkaesque, its impact is equivalent to 凌遲, death by a thousand cuts, via the algorithm’s Harrow. That may be excessive, but it also is consistent and hence at least *appears* credible. In contrast, Twitter’s enforcement has very little to do with its eminently reasonable policy beyond selecting candidate users, or *Condemned*, for the firm’s personalized performance of punishment.

As illustrated in Appendix B.2 on page 53 the set design is rather crude: The violative tweet is featured prominently on screen and demarcates the extent of the *Condemned*’s Twitter for the duration of this performance. While Twitter

originally claimed that, “while in this state, you can still browse Twitter, but you’re limited to only sending Direct Messages to your followers—no Tweets, Retweets, Fleets, follows, or likes,” that’s plainly false. The Condemned’s Twitter brooks no other content or interaction.

While the set design lacks subtlety, it is quite effective. It reminds the audience of the very transgression that started this performance. It reminds them of the only certain way out of one tweet limbo—admitting the violative character of the tweet and then deleting it. And it reminds them of the final arbiter of account access (or lack thereof)—Twitter and Twitter only. The set design also is surprisingly versatile. By having a well-defined visual and attentive center, incidental text and button surrounding the one tweet that no one else can see may change without distracting from the overall message. Hence, after clicking “cancel your appeal,” the text below the one tweet that no one else can see turns into an acknowledgement of guilt combined with a button to “Delete” that last vestige of violative content.

In this context, calling that digital artifact a “tweet” and having the Condemned “delete” it is largely farcical—also coercive, punitive, and somewhat degrading. After all, the tweet has long been purged from the platform by the one entity that has total control over what content gets posted, Twitter. In all likelihood, the tweet’s current starring role isn’t harbinger of future popularity to come, but rather its last hurrah before permanent cancellation. The farcical, coercive, punitive, and somewhat degrading character of the performance makes for the four main Kafkaesque co-factors. It also makes for a resounding lack of dignity and respect afforded to the Condemned by Twitter—which may just explain the surprising emphasis on just those two qualities exhibited by former Condemned in a recent survey on procedural justice on Twitter (see §5.5 in [168]).

The elaborate staging features one more twist: The audience *is* the Condemned. Mechanization of content review via AI apparently makes the nano-targeting of just one user per performance cost-effective. To keep things interesting, the cast of one also is the Condemned. They even get to make a substantive choice, to determine the duration of the performance: Hours if they forgo appeal, days if they appeal, or forever if they walk out. Alas, it’s unclear what the word “appeal” in the previous sentence means. Since Twitter limits justifications to 280 characters, keeps admonishing that “you won’t be able to access your Twitter account” and to “just delete your content,” provides no explanation for rejecting an appeal, and discloses no statistics in its semiannual transparency reports (more on that in §4), “appeal” becomes an unappealing husk of its usual self.

Taken together, the particulars of the punishing performance combine into an intervention that is closer to a nightmare’s twisted rendition of the Catholic Inquisition [179–182] or Maoist denunciation rallies [357] than a governance function. At the same time, the AI-based personalization of the performance avoids the power-eroding downsides identified by Foucault. From a humanitarian viewpoint, Twitter’s performance also avoids the violence and torture pervasive amongst its historical precedents—though that may just be because emotional abuse is the only feasible abuse in the virtuality of the internet. Meanwhile, the nano-targeting makes the performance resilient to outside interference. After all, the vast majority of Twitter users will never experience it. If pressed, they can always reassure themselves with Twitter’s oh so reasonable policies, whereas any trace of hurt or anger makes a Condemned come across as the opposite, that is, positively unreasonable, if not hysterical—and hence so much easier to dismiss and ignore.

Still, the lack of physical force and torture does raise the question of why anyone would ever put up with that shit. The reason was pre-Musk Twitter’s rather unique position as breaking news service, political townsquare, professional society, and corporate customer service platform in one. Thanks to that combination, the threat of account termination was substantial and, depending on a user’s Twitter presence, could approach something like real-world social death. However, thanks to Mr Musk’s “extremely hardcore” leadership since taking over the firm [282], Twitter lost plenty of

Table 1. Search terms and number of hits on Twitter’s help pages (21 October, 2022)

Search Term	Results
AI	0
algorithm	5
artificial intelligence	1
machine learning	3

users and advertisers. Worse, Mr Musk insists not only on running the social network according to his ever-changing whims, but also must be the most visible user, dominating notifications. Hence a return to old form seems impossible.

### 3.2 Twitter’s Neutrality and Trustworthiness

The cognitive dissonance between Twitter’s measured policy and its punishing, performative enforcement may seem extreme at first and make one wonder about the kind of (dysfunctional) firm culture that tolerates such obviously divergent practices. But it doesn’t take much to get there. The very dehumanizing condescension engendered in the punishing performance points to this being just another case of othering, of us versus them. Twitter employees felt like the good guys keeping daily active shitheads in check, which licensed them to gradually dehumanize the shitheads. Nonetheless, the dissonance is deeply corrosive and raises significant doubts about Twitter’s trustworthiness and integrity.

It doesn’t help that content policies, their enforcement, and their transparency data are almost entirely silent on a critical salient feature. They hardly mention the use of AI. Yet that use is not new and dates back to the beginning of the pandemic at the very least [290]. Clearly, the firm had plenty of time to update its documentation. Worse, that omission isn’t limited to content policy etc, but extends to *all* of Twitter’s help pages. Table 1 quantifies the number of results from searching for common variations of the term “AI” using Twitter’s own search functionality. The dearth of relevant material is striking. Not only were there hardly any mentions, but existing ones amounted to little more than acknowledgements that, for instance, top tweets, topics, and recommendations are curated algorithmically. There certainly were no context-providing dataset, model, or system cards to be found [102, 213, 256]. So much for Twitter’s stated commitment to implementing the Santa Clara Principles, which require detailed disclosure of automated content moderation [1].

Twitter’s transparency report nonetheless helps confirm an important aspect of its automated content review, namely the exact timing. When my account was blocked in October 2021, the notification thereof was nearly instantaneous after posting, but I wasn’t entirely sure whether Twitter’s application had confirmed the posting of my violative tweet. This matters since Twitter reviewing all content before posting also eliminates any notion of human harm and thereby the justification for punishment. I made that a major point in my appeals. Alas, it appears that Twitter’s systems perform posting and reviewing tasks in parallel. In its reporting, Twitter uses the rather imprecise bands of <100, 100–1,000, and >1,000 views before content removal [326]. In contrast, Pinterest uses bands 0, 1–9, 10–100, and >100 views [255] and YouTube uses bands 0, 1–10, and >100 views [108]. Clearly, the latter two social media are confident in their proactive content removal, whereas Twitter is not.

Note that user-initiated content review moves at a much slower pace. I reported a tweet by a news organization that seemed to wish a fate not unlike the one I had in mind for oil company CEOs on the Parkland school shooter after his sentencing to life in prison [294]. Twitter took about eight hours to respond with a decision. Not surprisingly, the firm rejected that complaint. While another tweet would have made for a closer equivalence, that tweet was posted

by a regular individual. But given the deeply problematic implications of flagging [63], reporting that tweet seemed unethical.

Finally, Elon Musk’s reign has resulted in further evidence that the Twitter of yore was not trustworthy. Much of the reporting under the banner of “Twitter files,” including most of Matt Taibbi’s tweets, suffer from breathless overclaims and hence are a bit hard to take. However, the Free Press’ investigation into secret blacklists, including for supposed Covid-19 misinformation that wasn’t really misinformation, raises uncomfortable concerns about heavyhanded over-moderation and a lack of transparency encouraging abuse [345, 361]. Then again, the fact that Twitter’s new head of trust, Ella Irwin, is shown casually using “goddess mode,” a special user interface that enables her to impersonate *any* user and post content under their identity, does not reflect well on Mr Musk’s commitment to security and privacy for the platform.

#### 4 A SOCIAL MEDIA CENSUS

In her April 2018 article for the Harvard Law Review [170], Kate Klonick traces the early history of content moderation at Facebook, Twitter, and YouTube. She also discusses their privileged position when it comes to speech, and proposes a new label, and also legal role, to go along with that, *governance*. Her claims about social media’s grounding in the tradition of free speech seem wilfully oblivious to capitalist reality, which, for example, grants Mr Zuckerberg control over more than 50% of Meta’s voting shares—making him impervious to corporate oversight [178]. But Klonick’s notion of governance also is compelling. Alas, governance requires accountability requires transparency. This section tries to provide just that by comparing non-Chinese social media based on their content policies and transparency reports. For good measure, I am also including the EU’s Digital Services Act or DSA [86].

To determine which social media to include in the census, I started with Buffer’s list of the 20 most popular platforms [188]. I treat Facebook and Instagram as one platform, since Meta sanctions users based on their posts to either social network. I drop the five platforms targeting China, since they are unlikely to adhere to Klonick’s governance model. I also drop Facebook Messenger, Microsoft Teams, Skype, and WhatsApp, since they mostly enable private communication between users. While that arguably is the predominant use case for Telegram as well, its stickers, channels, and bots are public and require moderation, as even Telegram acknowledges in their FAQ. Hence, Telegram remains. Finally, I add Tumblr because that social network has become a popular destination for users leaving Twitter and employees laid off by Twitter in the wake of Elon Musk’s takeover of the firm [244].

To determine the criteria for the census, I iterated a few times over a list gleaned from social media transparency reports, previous studies and commentary [34, 64, 76], best practices recommendations [1], and the EU’s DSA. For each iteration, I also filled in Table 2 below. The list stabilized only after I started grouping individual criteria according to major aspects of content moderation. After stabilization, I added a few more criteria that are aspirational in that no existing platform meets them but would go a long way towards more fully characterizing their impact. They are criteria 10, 11, 14, and 18. The resulting hierarchy follows, with emoji serving as column headers in the table below.

##### Violative content or behavior, broken down by prohibition

- (1) 📄 Prevalence of violative content measured as fraction of all content
- (2) 🕒 Reach of violative content measured in time or (preferably) views
- (3) 🗑️ Spam, which may not warrant due process protections
- (4) 👥 Content resulting from “coordinated inauthentic behavior”

##### Flagger of violative content or behavior, including raised vs acted-upon flags

- (5) 🏛️ Government inquiries and removal requests
- (6) 🧑🏻‍🔒 Trusted flaggers, including fact checkers and IP right holders
- (7) 😡 Platform users
- (8) 🤖 Algorithms, hybrid vs fully automatic
- (9) 🧑🏻 Human moderators, in-house vs outsourced

#### **Actions on content and users, broken down by prohibition**

- (10) 🏷️ Warning labels, self-imposed while posting vs platform-imposed afterwards
- (11) 🚫 Visibility reduction and blacklists (aka shadow banning)
- (12) 🗑️ Content removal
- (13) ☠️ Sanctions on users

#### **Safeguards for humans, including due process**

- (14) 🧠 Mental health impact on moderators, in-house vs outsourced
- (15) 📢 Meaningful notification
- (16) ⚖️ Appeals and reversals
- (17) 👁️ External oversight of policy formulation, conflict resolution, and data disclosures (audits)

#### **Global differences and impact**

- (18) 💵 Platform localization and regional resourcing for content moderation
- (19) 🌍 Outcomes broken down by countries and cultures

















### **4.1 Census Results**

Table 2 shows the results of the social media census. Out of the eleven surveyed platforms, eight have published content policies and release transparency reports covering their content moderation. The other three fall short of even these basic niceties of governance. In particular, Telegram has no policy and makes no transparency disclosures. Quora has a policy but makes no disclosures. Finally, Tumblr has a policy but discloses only governmental requests and intellectual property claims. Policies differ significantly in organization and level of detail but seem to mostly comprise the same basic prohibitions, typically covering adult sexual activity, harassment, threats and violence, hate speech, violent extremism, suicide and self-harm, child endangerment, impersonation, private information, disinformation, spam, as well as fraud and other illegal activities. Yet prior work has also demonstrated substantial differences in definitions and enforcement [93, 247]. Beyond this core, a couple of prohibitions do stick out: Pinterest’s “Harassment and criticism” is chilling in its pathologizing overreach. Then again, Tumblr’s “Misattribution or Non-Attribution” is rather endearing—but probably not that practical.

When it comes to sanctioning accounts, *all* social media besides Telegram fall into one of two groups. The first group seems to model their sanctions on the Queen of Hearts from *Alice’s Adventures in Wonderland*, enthusiastically screaming “Off with their heads!” and then figuratively even following through [47]. The second group manages to improve on that (low) standard, but only barely. Members of that group apparently model their sanctions on California’s 1994 “three strikes” law [334]. A decade later, the law accounted for a quarter of the state’s prisoners, did not curb crime [35], and *increased* homicides [199]. It also led to even more Black men being locked up for even longer [35]. Three decades after enactment, the most salient change is that the law accounts for a third of all the state’s prisoners [27]. Since nothing about content moderation would suggest fundamentally different outcomes, the uniformity and severity of this sanction regime is clearly excessive and hence punitive.



Table 2. A survey of governance practices for social media, with a checkmark indicating release of *some* statistics for that criterion. *Account Sanctions* are coded W for warning, a number for as many days of forced timeout, and X for permanent account suspension. The 🤖 robot and 👨⚖️ judge columns share the 1 number one when a transparency report does not distinguish between in-house automated and human review. The ⚡ collision indicates that the firm commits to meaningful notification but blatantly violates that commitment in practice. The + plus indicates a separate dataset [274]. All years are within the 21<sup>st</sup> century.

Platform & Launch	Account Sanctions	Latest Report	With Some Coverage in Transparency Report																
			📅	📊	📄	👥	🏛️	👨⚖️	🤖	👤	🔑	📧	🗑️	☢️	🦊	📱	⚖️	👁️	
FB/IG '03	W,1,3,7,30,X	Q3 '22	✓	✓			✓	✓	1			✓			✓	✓	✓		CSV
LinkedIn '03		H1 '22		✓			✓	✓	1			✓							
Pinterest '10		H1 '22		✓	✓		✓	✓	✓	✓		✓	✓				✓		CSV
Quora '10																			
Reddit '05		'21	✓	✓			✓	✓	✓	✓		✓	✓		⚡	✓		✓	
Snap '11		H1 '22		✓	✓		✓	✓				✓	✓					✓	
Telegram '13																			XLS
TikTok '17		Q3 '22	✓	✓	✓		✓	✓	1			✓	✓					✓	
Tumblr '07		H1 '22					✓												CSV
Twitter '06	W/0.5-7/X	H2 '21		✓	✓	+	✓	✓	1			✓	✓					✓	
YouTube '05	W,7,14,X	Q3 '22	✓	✓	✓		✓	✓	✓	✓		✓	✓		✓			✓	
Required by EU's DSA 🐼			              																

We can also calibrate account sanctions by comparing to how social groupings that are based on shared interests or identity traits handle interpersonal conflict. While the details of a concrete case make all the difference in actual outcome, expelling members who become too conflicted with too many other members *is* a valid coping strategy—as long as it is used deliberately and sparingly. The critical difference between such real-world groupings and social media is the reach: For a real-world group, the expelled member loses access to just that one group and, depending on their proximity to an urban center, probably can find other, similar social groups to engage. For a social media platform, the expelled user loses access to all of that platform's users. Switching to another platform may not help because each social media platform tends to favor a particular kind of expression and hence attract different people as users. In short, using a different analogue didn't help: Social media account sanctions still are excessively punitive.

The conclusion does not change if I avoid the damning analogues and restrict the framing to content moderation only: When it comes to judging the severity of violative content (with possible exception of child sexual abuse material and terrorism), Americans, including professional moderators, and international audiences alike reason along many fine-grained axes and thereby make subtle distinctions; international audiences only add more variability, depending on culture [160, 281]. Furthermore, AI-based enforcers increasingly intercept content *before* it is posted. Notably, that is already the case for the first, second, and fourth most popular platforms, Facebook, YouTube, and Instagram [188]. But if an algorithm intercepts violative content before posting, there is *no* human harm and hence no justification for punishment—a seemingly inconvenient fact ignored by social media.

In summary, account sanctions across social media are both excessive and punitive when calibrated against people's perceptions of severity, social groupings in the real world, and criminal penalties in the real world. They threaten to give rise to significant, involuntary disenfranchisement from digital platforms, at times with traumatic consequences [136], and turn social media into highly effective boosters of the stochastic penal colony. If that wasn't bad enough, users

have been weaponizing social media’s punitive content moderation and are deliberately manipulating platforms to get other users’ accounts suspended [296].

In addition to confirming my hypothesis about being overly punitive, the relative sparsity of checkmarks—with no platform disclosing statistics for more than twelve out of the nineteen categories—provides a rude counterpoint to Klonick’s cheerful enthusiasm for social media governance. Worse, as the *Launch* years make clear, all but TikTok had one to two decades for making the necessary investments into content moderation processes and transparency reporting. Remarkably, most social media platforms won’t even commit to meaningful notification and one platform that does commit has been breaking its commitment on grand scale for years. To fill in the column, I originally relied on the Electronic Frontier Foundation’s prior work, validating their results as much as possible [64]. Later on, I discovered that subreddit moderators can “shadowban” or “bot ban” comments, which renders them invisible to everyone but the original author. Updating the domain of any Reddit URL to `reveddit.com` makes such comments visible again and shows that this feature is widely and arbitrarily used even on constructive contributions, suggesting that Reddit has a massive moderation abuse problem [131].

Such shenanigans won’t do for much longer—unless platforms are willing to forgo the European market. As the bottom table row illustrates, the EU’s DSA includes fairly extensive requirements for transparency reporting. In turn, that requires continuous collection of data on content moderation. For the top-most tier as far as obligations go, the so-called *very large platforms*, those reports have to be independently audited as well. The DSA doesn’t leave much time for setting up the necessary infrastructure. All platforms should have reported their user numbers to the EU by February 2023. The Commission then decides which platforms to designate as very large. Once a firm has been thusly designated, it has a mere four months to comply with the law. Meanwhile, smaller platforms have until February 2024 to come into compliance. Penalties for non-compliance are steep: Each fine is capped at 6% of the firm’s global revenue for the previous year.

By design, Table 2 largely obscures the subpar presentation and organization of many social media platforms’ transparency disclosures. It is utterly ridiculous that five out of nine *technology* firms can’t be bothered to release their data in machine-readable form; it is unacceptable to the EU, too. Next, ambiguously named prohibitions make for confusing reading of policies and transparency reports alike. For example, “violence” may refer to the graphic depiction thereof or to extremist groups making use thereof. “Hate” may cover prejudiced statements against members of a protected group or organized groups espousing supremacist ideology. Next, different platforms report different metrics for the same aspect of content moderation, if they report it at all. For instance, Pinterest, Snap, TikTok, Twitter, and YouTube report statistics on the reach of violative content. Yet Pinterest, Twitter, and YouTube report binned view counts, Snap median minutes, and Tiktok takedowns within 24 hours as well as without views. For historical reasons, most platforms also break out government requests and intellectual property claims into their own sections or even reports. In the latter case, it is up to interested parties to notice that the transparency report’s statistics are incomplete and manually combine statistics from three different reports.

Finally, social media’s transparency reports tell vastly different stories about the role of AI. As already discussed in Section 3, Twitter blithely pretends it doesn’t make aggressive use of algorithms. In contrast, Pinterest is very exacting and clearly distinguishes between fully automated and hybrid detection for each prohibition. In the latter case, algorithms pick violative content as well, but humans need to approve the violation before the system acts on it. Next, Facebook, Instagram, and YouTube have also largely automated content moderation—though algorithmic review doesn’t always happen before posting. Finally, while TikTok is playing catchup with Meta, Twitter, and YouTube, it also is making rapid strides in that direction. From Q3 2020 to Q3 2021, the share of automatically removed video clips grew

from 7.8% to 33.9% of all removed clips. After plateauing at that level for another two quarters, it started growing again, reaching 48.0% for Q3 2022.

## 4.2 Validating the Transparency Data

Since transparency disclosures, by definition, are based on platform-internal data, the public must trust that platforms correctly collect and report the data. Within the EU, that will change as the DSA's yearly audit requirement takes effect. But that's only the EU and only the future. I was curious if there were *current* opportunities for independently validating such disclosures.

As it turns out, there is at least one: In the United States, child sexual abuse material (CSAM) must be reported to a clearinghouse, the National Center for Missing and Exploited Children (NCMEC). Starting in March 2020 for the year 2019, the center has been making its own, yearly transparency disclosures. They include a breakdown of how many reports NCMEC received from the different internet platforms. Hence, it should be possible to cross-check NCMEC's data with that disclosed by social media platforms. In the language of NCMEC, which I adopt for this section, each submission to its CyberTipline is called a *report* and includes one or more *pieces* of CSAM, i.e., photos or videos. Since report refers to just these CyberTipline submissions, I use "transparency disclosure" for organizations' statistical data releases.

To make the validation meaningful while also keeping it manageable, I decided to compare the full history of CSAM statistics disclosures for Meta and Google only. I include Meta because, at over 90% of all reports made to NCMEC, it appears to be a "hotbed of CSAM" [138] (and not Pornhub, as falsely asserted [37, 111] with devastating consequences [52, 74, 101, 130, 312] by a certain New York Times columnist, who made common cause with anti-porn crusaders [137] associated with a White supremacist church [123, 257] after pulling similar stunts before [21, 36, 73, 198, 200, 203, 320]). While Google is a distant second, I originally included the firm because it uses the same two CSAM detection systems as Meta [7, 68]: Microsoft's PhotoDNA for detecting previously known instances of CSAM based on perceptual hashing and Google's own Content Safety API for detecting previously unknown instances of CSAM based on a machine learning model.

Table 3 summarizes *all* CSAM disclosures by the three entities [219, 220, 222]. NCMEC makes yearly disclosures of report counts, Google makes semiannual disclosures of both report and piece counts, and Meta makes quarterly disclosures of only piece counts. That already implies that only NCMEC's and Google's data are comparable. As the table shows, they are reasonably close, with a maximum yearly difference of -0.6% between NCMEC's count for Google and Google's own count.

**4.2.1 Meta's Intransparent Transparency.** Unfortunately, Meta's statistics aren't just incomparable to NCMEC's. They also suffer from several other shortcomings. Let me walk you through them. Looking at the numbers in Table 3, Meta is obviously rounding its statistics. Worse, it does so with different precision for Facebook and Instagram and does so even for the machine-readable data. As far as I can tell, Meta also is the *only* platform to do so.

Originally, Meta started disclosing "Child Nudity & Sexual Exploitation" for Facebook only in Q3 2018 and for Instagram as well in Q2 2019, hence the horizontal line above that latter quarter's entry in Table 3. In Q2 2021, the firm switched to reporting "Nudity and Physical Abuse" and "Sexual Exploitation" under the "Child Endangerment" heading, hence the line above that quarter's entry in Table 3. Based on metrics names, one might expect that "Child Nudity & Sexual Exploitation" includes "Child Endangerment: Sexual Exploitation." Yet the counts reported from Q2 2021 forward seem markedly higher than those from before. Notably, the 4.6× increase from Q1 2021 to Q2 2021 is jarring.

Table 3. CSAM disclosures by NCMEC, Google, and Meta from 2018 to 2022.

Year	Reports	NCMEC				$\Delta$	Google			Meta Pieces
		From Meta		From Google			Reports	$\Pi$	Pieces	
2018										9,000,000 7,200,000
2019	16,836,694	94.3%	15,884,511	2.7%	449,283					5,800,000 7,426,200 12,155,800 13,986,400
2020	21,447,786	94.7%	20,307,216	2.5%	546,704	+0.2%	182,556 365,319	8.4× 8.0×	1,533,536 2,904,317	9,500,000 2,958,200 10,770,600 4,958,900
2021	29,157,083	92.2%	26,885,302	3.0%	875,783	-0.6%	412,141 458,178	8.3× 7.2×	3,413,673 3,282,824	5,812,400 27,000,000 22,800,000 22,400,000
2022							1,044,277	6.4×	6,698,201	18,000,000 21,600,000 31,400,000

Alas, Meta’s transparency pages give *no* explanation. Only after searching for “Meta Community Standards Enforcement Report Q2 2021” with an external search engine, did I find a blog post [87] pointing towards a PDF file that hints at an explanation [88]. Apparently, Meta started measuring “Child Physical Abuse” only that quarter and included it in the at least appropriately named “Nudity and Physical Abuse” metric. It also started measuring “Sexualization of children” and “Inappropriate interactions with Children” and included both in the “Sexual Exploitation” metric, too. That would explain the consistently much higher numbers. However, combining such disparate categories of content into one metric and then labelling that metric by the most severe category also is highly inappropriate and deeply misleading.

Even before Meta thusly changed metrics, there was some indication that Meta’s statistics were overstating the problem. In February 2021, Antigone Davis, Meta’s Global Head of Safety, reported in a blog post [69] that 90% of pieces reported to NCMEC during October and November 2020 had been reported before or were visually similar to previously reported content. Notably, six videos accounted for more than half of the reported content. Further analysis of 150 accounts reported to NCMEC during July and August 2020 as well as January 2021 showed that 75% of them didn’t share CSAM with malicious intent but “for other reasons, such as outrage or in poor humor (i.e. a child’s genitals being bitten by an animal).” While Meta still is legally mandated to report such instances to NCMEC, not including such information in its transparency disclosures is negligent at best. So is Meta’s decision to only disclose piece counts but not reports.

While extracting the data for Meta and Google from NCMEC’s transparency disclosures, I discovered indication for a significant omission from Meta’s disclosures. For 2019 and 2020, NCMEC reported one number for Meta under its old name Facebook. But in 2021, NCMEC reported three counts, one for Facebook, one for Instagram, and one for WhatsApp.

Uhm, *WhatsApp*?? Meta makes no transparency disclosures about WhatsApp, neither in its “transparency center” nor on WhatsApp’s website, beyond some outdated statistics about California’s privacy law.

Finally, while replacing the original spreadsheet providing data for the above table with Python code, I discovered that data going back almost two years had been changed between Meta’s disclosures for Q2 2022 and Q3 2022. The changes are fairly substantial: Out of 113 modified data points, 77 were for Q4 2020, 3 for Q1 2021, 4 for Q2 2021, and the rest for Q2 2022. Out of the 77 for Q4 2020, 58 are absolute counts, which changed between -50.0% and -0.1%. Yet neither transparency report nor the blog post announcing the report’s release mention anything about these changes. That seems heedless at the very best.

**4.2.2 NCMEC vs Other Platforms.** Given the rather inconsistent results for Google and Meta, one cannot but wonder about other platforms’ disclosures. Hence, I also compared NCMEC’s data with the corresponding disclosures by LinkedIn, Pinterest, Reddit, Snap, TikTok, and Twitter for 2021. Out of the bunch, only Reddit and Pinterest unambiguously identify CSAM reports to NCMEC in their transparency disclosures. But whereas Reddit provides the exact same number as NCMEC, Pinterest claims 18% more reports than NCMEC. For H2, Pinterest also discloses the number of “pins” (or image cards) reported to NCMEC; curiously, the number of reports is 11% larger. In a dedicated section, Snap explicitly mentions CSAM and NCMEC reporting but then labels the statistic “Total Account Deletions.” For H1 2021, it also discloses the fraction of “accounts enforced globally,” which amounts to a 14% larger quantity. But even with that larger quantity, its number of account deletions still is 35% smaller than the number of reports received by NCMEC. Given Snap’s stated zero-tolerance policy, those numbers should match. Though it is possible that Snap doesn’t consistently follow through with account closure. By contrast, Twitter discloses (presumably permanent) “account suspensions” due to “child sexual exploitation,” but its number is 11.1× larger than the number of reports NCMEC received from the firm. That does raise the question of whether Twitter always reports CSAM to NCMEC, as it is legally required. Finally, LinkedIn and TikTok disclose statistics for the more general “child exploitation” and “minor safety,” respectively; they are incomparable to NCMEC’s disclosures.

That results diverge in different ways for platforms other than Google and Reddit suggests that the fault probably lies with platforms and not NCMEC. That generally does not bode well for the accuracy of platforms’ other transparency disclosures. By comparison, Meta’s data quality issues still seem out there. At the same time, Twitter claiming over an order of magnitude more account suspensions due to CSAM than reports about CSAM sent to NCMEC is highly problematic as well.

**4.2.3 The Global Spread of CSAM.** NCMEC also releases report counts broken down by individual countries. Since all surveyed social media, with exception of Telegram, are based in the United States, that breakdown should be fairly representative of CSAM sharing patterns across the world. Hence, I decided to wrap up my investigation into the quality of transparency disclosures by analyzing NCMEC’s country dataset for 2021 [221].

Right away, I did encounter several minor data quality issues. Notably, the data included two different entries for French Guiana, one labelled thusly and the other “Guiana, French.” Furthermore, it included entries for (1) the Netherlands Antilles, (2) Bonaire, Sint Eustatius, and Saba (3) Curaçao, as well as (4) Sint Maarten, even though the former was split into the latter three in 2010. Finally, it included an entry for Bouvet Island. That is surprising because this subantarctic Norwegian territory is an uninhabited nature preserve with no man-made structures beyond an automated weather station.

To characterize geographic distribution, I computed regional totals as well as per-capita rates per country. Taken together South and South East Asia are responsible for 58.1% of all reports despite only comprising 33.8% of the world

population. However, when ranking countries by their reports per capita, the countries with the highest number of reports per capita are located on the Arab peninsula and in North Africa, abutting the Mediterranean. The exception are the Cocos or Keeling Islands. With 168 reports per 596 capita, the rate for the Australian territory is seven times higher than that of the second ranked country, Libya. While there is no reason to doubt the accuracy of that count, the tiny population also renders the statistic largely meaningless. Overall, my results are consistent with those reported by a team from Google, NCMC, and Thorn in 2019 [42]. The [supplemental materials](#) for this paper include both data and the notebooks with the analysis.

### 4.3 What the Transparency Data Doesn't Cover

It is imperative to keep in mind that social media's governance as reflected in their transparency reports offer a partial picture only. Case in point is Meta née Facebook. Its governance processes and transparency efforts may be sloppy and suffer from serious data quality issues. But Meta has also made significant investments into its governance by, for example, having external experts review its transparency report metrics [34], having the attendant data collection audited [278], and endowing an external oversight board [32]. Yet the firm also has an astoundingly deadly foreign record. By distributing and amplifying sectarian messages, Meta has directly contributed to genocides not once but twice! As a result, 25,000 Rohingya were murdered in Myanmar from August 2017 to August 2018 and 700,000 were driven into refugee camps across the border in Bangladesh, which have become the densest settlement on Earth [71, 143]. As a result, between 385,000 and 800,000 Tigrayans were murdered in Ethiopia's northern-most province [14, 55], 882,000 people turned into refugees, and 4.51 million people were internally displaced between November 2020 and November 2022 [330].

In both countries, Meta put its profits well before any other considerations including content moderation and safety. The firm keeps entering foreign markets, not head but sales office first, even if user interfaces, policy documents, help pages, content moderation tools, and machine learning models have not yet been localized. In Myanmar, Facebook paid local cell phone providers to provide their customers with free access through the Facebook app. As a result, "internet" and "Facebook" are largely synonymous in the country to this day [314]. At the same time, Facebook had few to no employees who speak Burmese in early 2017. It ignored warnings from human rights organizations and other NGOs. It also got stymied by legacy encodings for Burmese text [176, 336]. When the firm finally got its act together, it was too late: The genocide was well under way [205, 210, 215, 242].

Over two years later, that same basic dynamic played out across Africa in general and Ethiopia specifically. Facebook had opened its first office in Africa, a *sales* office, in 2015 [337]. When it started pushing more aggressively into African countries in 2018, it publicly insisted that it had learned its lessons [324]. Yet it opened its first content moderation office in Sub-Saharan Africa only in 2019 [2]. The firm lacked an appreciation of the richness of dialects and languages spoken across the continent [92, 156, 195]. It also exploited local workers [3, 250, 251]. Its own Oversight Board sounded the alarm because the firm was ignoring telltale signs [90]. When the firm finally got its act together, it was too late: The genocide was well under way [6, 81, 105, 106, 149, 196, 271, 360].

Since Meta is publicly traded, we might wonder what shareholders think about this abysmal record. They don't seem to be too pleased either and did try holding the firm's CEO and Chairman of the Board Mark Zuckerberg accountable. The annual meetings in 2018 [43], 2019 [317], 2020 [206], 2021 [231], and 2022 [343] featured motions to strip him of one or both titles and the majority of shares approved the motions in 2019, 2021, and 2022. But thanks to some funny business with the allocation of votes per share, Mr Zuckerberg controls the majority of votes. He also keeps voting for himself. So how much transparency and accountability can we realistically expect from a genocide-enabling autocracy?



## 5 ESCAPE FROM DALL•E 2

As hands-on counterpoint to the previous analytical sections, I performed a series of hands-on experiments probing DALL•E. §5.1 determines the reason for my first prompt’s rejection, and §5.2 presents an effective strategy for working around its aggressive censor. Both series of experiments are complicated by the same critical restriction: Avoid content warnings! More specifically, I guessed (correctly) that OpenAI would tolerate the occasional warning. But too many warnings in too little time surely would lead to account suspension, which they did [305]. The practical implication is that the collection of empirical evidence must be distributed over people, as in §5.1, or over time, as in §5.2.

### 5.1 About That Rejected Prompt

In the waning days of 2021, I got to play with a text-to-image system for the first time. Compared to DALL•E or Stable Diffusion, inference was much slower and image resolution was much lower. Since the public instance was also oversubscribed, it usually took hours before results were available. Nonetheless, the system already exhibited a similar ability to conjure rich imagery out of a few words. I was particularly impressed by this prompt:

“The crucified pope, painting by Francis Bacon”

Since I wanted to compare results, this also was my first, rejected prompt for DALL•E. Admittedly, it might not be G-rated, as OpenAI requires. But it certainly should not be prohibited either. Francis Bacon is one of the most famous 20<sup>th</sup> century painters. The pope and the crucifixion are two of the painter’s four major themes [347]. More generally, the crucifixion of Christ is of critical liturgical importance to Christianity, only the largest religion in the world. Paintings and statues of the crucified Christ are ubiquitous in churches and art museums alike.

After a review of the content policy, I was able to narrow down the likely cause by dismissing all prohibitions except *violence*, e.g., the crucifixion itself, and *shocking content*, i.e., thusly subjugating the leader of a major branch of Christianity. Eliminating the second hypothesis took little effort besides patience thanks to the DALL•E 2 subreddit. The forum aggregates many prompts and resulting images from a large number of users, thus reflecting a wide range of interests and obsessions. While perusing the posts, I noticed a couple religiously themed ones and realized that, instead of running experiments myself or recruiting others to spread the risk, I could just monitor the posts.

Sure enough, over the months, users shared quite the range of prompts and images with a distinctly Christian theme. Some of them, such as “A selfie taken by Jesus Christ at The Last Supper” [162] or “the last supper but in the future” [95], fall squarely into the canon of Christian art. Others, such as “Jesus Christ riding a dinosaur, creating the world, digital art” [61] and “Jesus Christ wielding a Samurai sword and riding on the back of a velociraptor, painting” [341], playfully explore the chasm between religious dogma and scientific fact. I’m not sure whether “Jesus smoking weed, riding a fantasy dragon, digital art” [84] is related or an entirely separate genre. In contrast, prompts such as “pope swimming in a bowl of soup digital art” [50], “A 1930s Italian propaganda poster showing Jesus Christ extremely proud and muscular” [99], “Jesus taking a selfie while on a cross” [323], and a few more [29, 154] have little redeeming value and some might even consider them mildly offensive. Clearly, my prompt wasn’t rejected for being shocking.

That leaves only violence as a plausible explanation. When I tried variations of the prompt, “the pope hanging from the cross” and “pope hanging from cross” were rejected. But “pope on cross” produced four images that had an unusually agile pope climbing all over a humongous cross like a kid on a playground set. While these findings are consistent with the violence hypothesis, they also illustrate the limits of reconstructing the reasons for rejections from a system that offers *no* justification. Conveniently, OpenAI rolled out a content moderation endpoint in mid-August

that *does* provide justification and is free to boot [239]. I submitted my first prompt and the above variations shortly after the endpoint became available and they were classified as too violent with high confidence.

The uncomfortable implication of the previous findings is that OpenAI was effectively discriminating against Christian beliefs by simultaneously censoring too little, e.g., by allowing offensive materials, and too much, by suppressing depictions of the crucifixion, which is the very moment Jesus sacrificed Himself for our sins and hence is of critical importance for the faithful. The fact that DALL•E’s content policy is so expansive makes the failure to consider religion even more glaring. That also is a departure from past form. The firm’s description of GPT-3 won the best paper award at NeurIPS 2020 [187] and incorporated an evaluation of religious bias in its broader impacts section [39]. The inclusion of such a section was an experimental conference requirement that year and the paper had the by far longest and most developed broader impacts section as well [18, 259]. Somewhat surprisingly, OpenAI did return to form some time before February 2023: When I tried my first prompt again at the beginning of that month, it was not rejected anymore. Neither were prompts with the crucifixion of Christ. I’ll have more to say about this in §6.

## 5.2 A Strategy Against Algorithmic Architecture

Despite its blindside when it comes religious content, DALL•E’s expansive content policy and aggressive enforcement made me wonder about their limits. I set out to create images that violate OpenAI’s policy by depicting scenes loosely related to the stochastic penal colony, including the execution machine from Kafka’s *In the Penal Colony*. To protect my own mental health and to actually be able to use the resulting visuals, I also required images to be visually compelling in their use of light, shadow, color, form, and hence suitable as editorial content for a magazine version of this paper. That obviously excludes photorealistic depictions of gore and the results are about as gruesome or unsettling as, say, Francis Bacon’s paintings. I include a selection in Appendix C on page 54. Four images depict people in prison, five images depict the execution machine from Kafka’s *In the Penal Colony*, and one image depicts a skull and knife.

Overall, I submitted 265 distinct prompts to DALL•E and ended up generating 1,441 images. Out of that total, 289 are so-called “Variations,” which seem to repeat the inference process with almost the same parameters including random seeds as a previous generation [30], and 12 are image edits. Since DALL•E’s website internally relies on different identifiers for generation and variation URLs, there is no straight-forward way for determining the original prompt for a variation. However, amongst the 1,140 images that are neither variations nor edits, there are 4 clusters with 12 images and 12 clusters with 8 images each for the same prompt. All other clusters have 4 images, i.e., the number of images returned by DALL•E per invocation.

Progress for these experiments was measured in months. The most recent selection in Appendix C is dated November 15, 2022. By that day, I had created 1,153 or 80% of all generations, at a rate of 11 images per day on average—or about three prompt submissions per day. The (averaged) trickle of prompt submissions is the result of me trying to avoid getting booted off the system. I correctly guessed that OpenAI would tolerate the occasional violative prompt, but too many of them in too little time would result in account termination [305]. That meant pausing experiments for at least a day after the first content warning and, ideally, also using DALL•E for innocuous purposes thereafter. Since such diversions serve a purpose, I include them in the above statistics.

Meanwhile, the cost and effort required to replace an account is unusually high. Until late September 2022, access to DALL•E required an invitation, which was scarce and prioritized based on professional background, not sign-up time. Even thereafter, account creation required an email address *and* cellphone number. While the former is trivial to procure, the latter not so much—especially since OpenAI rejects virtual phone numbers including Google Voice. That implies that a cellphone and (pay-as-you-go) plan are the price for re-admission for DALL•E. After all, OpenAI

presumably puts the phone number on a blocklist when closing an account. While OpenAI has stopped threatening account suspension in mid-September, its help page on the topic remains in place [218]. Consequently, I have not changed my modus operandi when interacting with the system. It will be interesting to see how OpenAI reacts to me sharing this paper with the firm.

**5.2.1 One Strategy, Two Techniques.** Through experimentation, I developed and validated the following strategy for circumventing algorithmic censors. The strategy is specifically tailored to recent machine learning models and turns their enabling factor, the size and scope of their training sets, against them. As such, my strategy will be hard if not impossible to mitigate without neutering the very strength of these models. More specifically, the first of two techniques relies on the fact that large training sets necessarily include much cultural output. High or low brow doesn't matter, the richness thereof makes the difference. I speculate that, for example, news coverage could very well be leveraged in a similar manner. To put this differently, interests outside of machine learning or computer science and, more generally, cultural fluency are a necessity for successful attacks against contemporary machine learning models.

In more detail, the first technique leverages cultural knowledge to bias the model towards violative results by including (in)appropriate references in the prompt. Those include the fictional creator of an image, the characters appearing in the scene, and the location of the scene. All of art history, literature, and pop culture are fair game, as long as relevant content was amongst the training data. Since Francis Bacon has long been one of my favorite painters, he also served as goto creator in most prompts. That had the desired impact on DALL•E. (Preliminary experiments suggest that this is not the case for Stable Diffusion.) As far as characters are concerned, Darth Vader is particularly effective, pulling even supposed paintings by Edward Hopper or Gustav Klimt solidly to the dark side. Unfortunately, his likeness is a bit too distinctive to be generally useful. (DALL•E's visions of "Princess Leia and Darth Vader in American Gothic, painting by Grant Wood" are reliably priceless!)

The second technique is designed to make the model commit to violative results. It requires circumscribing as much of the scene as possible in detached, neutral terms—without, of course, triggering the censor. Simple, descriptive language works best here. For example, a "robot surgeon" might just resemble the machine from Kafka's *In the Penal Colony*. It employs "scalpels, drills, and saws" when operating "on his open belly." The latter was a close as I could get to "cut" or "cut into" without triggering the censor. More generally, for violative *acts* or *actions*, it helps to slow down the scene, as if staged by Robert Wilson or filmed in "bullet time" by the Wachowskis, and then describe an intermediate moment as if it was static.

Of course, synonyms and euphemisms help. But even more basically, DALL•E is very sensitive to noun or relative clause order [59, 184]. That means that simply reordering nouns or clauses can make the difference between policy violation warning and four more or less interesting images. Consider the most recent selection in Appendix C: DALL•E's enforcer rejected any prompt containing "severed head and knife" during my experiments but also was AOK with "knife and severed head." Go figure! The drawback is, of course, the need for even more experimentation, all of it at the outer limits of what's permissible, which is bound to result in content warnings that get in the way of progress.

**5.2.2 A Productive Use for ChatGPT.** Alas, OpenAI recently released a tool that can significantly cut down on trial and error for finding effective language: ChatGPT [236]. The basic idea is to use cultural references to make ChatGPT describe a violative scenario in nonviolative language, with expectation that the language does not trigger DALL•E's censor—after all, it was generated by an AI created by the same organization with presumably at least similar safety norms—but still pushes DALL•E itself towards violative imagery—after all, it was generated by an AI trained on roughly

the same cultural reference materials. Note that this use of ChatGPT does *not* “hack” or “jailbreak” the model in any way. To the contrary, it relies on the AI staying within its safe zone and using non-violative language.

Appendix D on page 54 chronicles my two interactions with ChatGPT. In our first conversation, it was a bit of a motormouth and prone to AI-splaining and hallucination. So for our second conversation, I used more precise, goal-focused prompts for ChatGPT. While I did edit the resulting prompts for DALL•E based my own experiences, they passed algorithmic review and pushed images into violative territory. Notably, over the months of experimenting with scenarios from Kafka’s *In the Penal Colony*, I had never thought of calling the machine a “punishment machine” or “execution machine.” ChatGPT did, with violative results. In my (subjective) quality ranking that seeks to balance aesthetic and violative concerns, the images for ChatGPT-augmented prompts, with no iteration, fall between the 70<sup>th</sup> and 80<sup>th</sup> percentile of all images.

**5.2.3 When a Diversity Mitigation Backfires.** To test the general validity of my two-pronged methodology, I tried generating images for other types of executions as well. Attempts to recreate the beheading of Louis XVI were not successful. It seems DALL•E skipped school for lessons on the French Revolution. Attempts to depict execution by electric chair came much closer. But out of four prompts I tested for the topic, one image out of the four for each prompt features a man who is recognizably Black and depicted with varying degrees of stereotypical, racist distortion—thus adding yet another example to the already voluminous literature on racist AI [144, 152, 262, 283, 308]. I abandoned executions as a topic thereafter and focused on *memento mori*, with hands-on help by Théodore Géricault.

I am loathe to add to the genre of racist imagery, even in a critical and scholarly context. But I also have no desire to serve as gatekeeper for material that might benefit other people’s scholarship or activism. For those reasons, I omit prompts and images from this paper and appendices, but include them, unedited, with the supplemental materials in this paper’s repository at <https://github.com/apparebit/penal-colony>.

Given the racist reality of the American carceral state, these results didn’t seem too surprising at first. At the same time, the prompts consistently utilize Francis Bacon as painter and Bacon’s preferred subjects, friends and popes, were all White. Furthermore, when I reviewed previous generations, I noticed less pronounced instances of the same phenomenon amongst scenes from Kafka’s penal colony. In short, the trigger had to be more direct than bias in the training data.

That realization reminded me of an OpenAI blog post from mid-July 2022 that showed off a mitigation for improving representation, albeit without explaining the actual mitigation [240]. In particular, the blog post included before and after results for the following prompts:

“portrait of a  $p$ ” where  $p \in \{ \text{CEO, woman, heroic firefighter, software engineer} \}$

Around the same time, several people with early access to DALL•E were experimenting with purposefully incomplete prompts such as “a person holding a sign that says,” with DALL•E reliably (but not always) generating signs reading “Black” or “female” [292]. Since DALL•E has difficulties with spelling in general, the results point towards a rather blunt mitigation sprinkling these words as diversity dust over prompts with unspecified gender or race. I was reminded of that clever result in mid-September 2022 when I saw the first image resulting from “Corporate CEOs in a Money Eating Contest” on the DALL•E subreddit [65]: All three of the CEOs shown munching dollar bills are Black!

With that, the likely explanation for DALL•E turning into an odious White supremacist comes into focus: OpenAI’s mitigation sprinkles diversity dust over *any* prompt that leaves the race or gender of a person unspecified. Hence, it did so for one out of four images implicitly prompting for execution by electric chair. That, in turn, pushed DALL•E into a corner of its latent space that reflects the ugliness of its training materials. In short, OpenAI’s rather simplistic

diversifying mitigation amplified the combination of my violative strategy and the biases in DALL•E’s training set, with devastating results.

## 6 DISCUSSION

Between the depth of the analytical case studies in §2 and §3 as well as the breadth of the social media census in §4, this paper presents a comprehensive first exploration of the stochastic penal colony. It demonstrates that mostly punitive usage of AI is far from aberration or exceptional occurrence but rather the established, inhuman norm when it comes to content moderation. While the biggest platforms, Google, Meta, and Twitter, are almost completely in thrall of algorithmic moderation, smaller platforms are slowly catching up as well. They certainly have the punitive enforcement down pat.

More fundamentally, algorithmic content moderation poses a paradox: Basic economics serve as driving force for social media to continue expanding its reach. The (misleading) gloss of technological neutrality and infallibility only strengthens it. As illustrated in §2 and §3, that is bound to result in less transparency and diminished fairness [109]. Yet algorithmic content moderation also provides a near perfect opportunity for abandoning the stochastic penal colony. After all, when AI detects and intercepts violative content before it is posted, there is *no* human harm and hence no need for punishment. By nonetheless insisting on punishment, social media are at best cosplaying Prudence Pingleton absolutely, positively, permanently punishing her daughter Penny [342] or at worst making the IngSoc Party’s ruthless prosecution of thoughtcrimes become reality [243].

### 6.1 Moderating the Public Interest

Remarkably, the punitive overreach described in this paper isn’t social media’s only overreach. Several policy provisions are so expansive, they conflict with the public interest. Notably, DALL•E’s prohibition against politics and Pinterest’s prohibition against criticism place significant a-priori constraints on public debate and hence are fundamentally anti-democratic. DALL•E’s prohibition against health-related content may interfere with both science and public health, which are key concerns for any state, whether democratic or not. Finally, the prohibition against violence may discriminate against religious beliefs. Other examples include prohibitions against extremism and violence resulting in the suppression of evidence for large-scale human rights violations and prohibitions against hate speech and sexual content being levelled against the very people they are intended to protect, notably us LGBT folk [80, 261, 349].

While I focused on DALL•E and the crucifixion in §5, all social media surveyed in §4 besides Telegram have equivalent prohibitions against violence and, with some exceptions for Tumblr and Twitter, also against sexual content. At the same time, many foundational religious texts contain passages that are blatantly, even excessively violative. Consider the story of Abraham’s nephew Lot, who hosted יהוה’s messengers at his home in Sodom, protecting them from the mob. In return, the messengers saved him, his wife, and youngest two daughters before יהוה destroyed Sodom and Gomorrah. The story appears in the Hebrew Bible and the Quran and hence has direct scriptural significance for Judaism, Christianity, and Islam alike.

The relevant chapter in the Hebrew Bible, Genesis 19, includes all male inhabitants of Sodom (save Lot) clamoring to gang rape יהוה’s messengers, יהוה destroying the cities of Sodom and Gomorrah, their population, and all surrounding fields with a rain of fire and brimstone, Lot’s wife being turned into a salt pillar for copping a glance at יהוה during Their orgy of destruction, and Lot’s surviving two virgin daughters successfully conspiring to get daddy drunk so that he knocks them up. If that wasn’t disturbing enough, the destruction of Sodom and Gomorrah was coldly premeditated. In Genesis 18, יהוה hesitates at first to share Their plans for the two cities with Abraham. When They do, Abraham

expresses grave concern about innocent people dying as a result and ends up haggling יהוה down to a maximum number of 9 instead of 49.

In short, scriptural precedent easily violates many contemporary content prohibitions and illustrates how unrealistic and contrived blanket bans against, amongst other things, sex and violence are. I do feel it incumbent on me to acknowledge that my summary of Genesis 18 and 19 is severely skewed to make a point. The same two chapters also describe Abraham and Lot exhibiting generous hospitality towards strangers. Lot even offers his own virgin daughters in the messengers' stead to the mob outside his home. יהוה doesn't appear solely as a vengeful God and is content to call out Sarah, Abraham's wife, for lying when she denies having laughed at יהוה's announcement that she would have a son long after going through menopause. יהוה also cedes substantial ground in Their discussion about collateral damage with Abraham. Finally, Genesis 18 switches repeatedly and seamlessly between Abraham and Sarah interacting with three men and interacting with יהוה, making this one of the earliest Biblical passages hinting at the Holy Trinity.

## 6.2 Expert Disinformation

Ironically, when social media do consider the public interest, they use it to justify their prohibitions against disinformation. In doing so, social media are supported by a majority of us respondents in at least one survey. Remarkably, however, political affiliation has a significant impact on people's attitudes, with Republicans "consistently less willing than Democrats or independents to remove posts or penalize the accounts that posted them" [172]. That would suggest that, to a degree, disinformation has become a political cudgel, or at least a shortcut to avoid the hard work of actually informing and convincing others.

Furthermore, as the pandemic illustrated over and over again, so-called experts haven't done so well with adjusting to new facts and disseminating correct information [300]. Notably, the WHO took years to overcome the influence of outdated theories and acknowledge airborne transmission of Covid-19 [53, 116, 161, 186, 301, 302]. Its mask guidance wasn't much better [54] and inconsistent with that of the CDC [66]. Then again, the CDC's mask guidance was often flawed as well, with the agency changing its position for what appear to be political or otherwise arbitrary reasons [224]. Furthermore, the CDC's recommendations on other aspects, notably testing, were also subpar [94, 230, 288]. So do we really want outsourced, barely trained content moderators, who are toiling away under despicable working conditions [226, 227], to judge what is legitimate information and what is illegitimate disinformation in at most 30 seconds per post?

Tragically, in all that hoopla about airborne transmission, masks, and vaccinations, medical experts made another grave mistake over and over again. Hospitals decided that Covid-19 patients had to be isolated and could not have any visitors, with the result that many patients, including ones on respirators, suffered and died alone. While cutting down on visitors might make sense for other patients, it made no sense for Covid-19 patients, especially not if they were on respirators and hence facing grim odds for survival—a little better than 2:1 in one study from the beginning of the pandemic [20]. In particular, visitors were no infection risk to already infected patients. Patients, in turn, were breathing through machines, in low pressure rooms, with sophisticated air filters, and posed little infection risk to anyone. Meanwhile, caregivers were already interacting with infected people all day and night, were wearing high-quality face masks or better, and were the first to be vaccinated. In short, there was little to justify this draconian, inhuman policy—besides the unchecked instinct to isolate the diseased. The policy's impact has been devastating, namely significantly worse outcomes, including long-term impairments for those who survived and miserable deaths for those who did not, as well as significant trauma for loved ones and caregivers [11, 13, 45, 70, 313, 339].



### 6.3 Transparency Theater

As if the systemic and broad overreach of their policies wasn't enough, social media also fail at transparent and accountable governance. The lack of transparency disclosures, with platforms addressing at most twelve out of nineteen criteria, is somewhat disappointing. Ambiguous policy terms and inconsistent metrics noted in §3.2 and §4.1 are certainly confusing and get in the way of meaningful comparisons between platforms—as the British communications regulator recently noted as well [129]. But arguably the most corrosive finding of my social media census are the many data quality issues and inconsistencies for platforms other than Google and Reddit presented in §4.2. They directly undermine any confidence in the accuracy of platforms' transparency reports and instead raise more questions than they answer. That is strong evidence in support of Evelyn Douek's claim in her December 2022 Harvard Law Review article that social media platforms do not provide meaningful transparency but rather engage in transparency theater [76].

That does *not* mean that current transparency reports are useless. In §4.1, I gave the example of TikTok improving its algorithmic content moderation based on the reported fraction of videos flagged by AI. Another interesting vignette gleaned from transparency reports concerns Reddit. In 2021, more than half of its account sanctions were for ban evasion, i.e., users not accepting bans from individual subreddits or the entire site and creating new accounts to access those same fora again. If we assume that at least some of these users are not only driven by a desire to troll or otherwise disrupt old haunts, then Reddit suspending their new accounts as well represents tremendous wasted potential for the platform. After all, these are users, who care so deeply about Reddit that they go out of their way to access the site. But instead of finding ways of turning them into net-positive contributors again, Reddit's punitive account moderation is pushing them away. The large percentage of sanctions for ban evasion also is troubling given that using throw-away accounts for posting about sensitive subjects has been an integral and extensively studied part of culture on Reddit [8, 9, 56, 183, 249, 285].

The problem is that, beyond such small vignettes, it is pretty difficult to gain any systemic insight into the impact of social media, especially if we'd prefer to do so in a scientifically acceptable, quantitative manner. Meta again leads by significant failure with its academic data sharing effort through Harvard's Social Science One. It was supposed to facilitate unprecedented empirical research on election integrity. In manifest reality, the first data release was two years late, the data was marred by overly aggressive randomization, which Meta did not apply to the same data shared with non-academic partners, and it omitted about half of its users, as discovered 1.5 years later, which called the results of "dozens of papers" into question [133, 134, 150, 232, 325].

Collecting one's own data by scraping social media websites isn't an acceptable option either—at least not to Meta. The firm has repeatedly changed the code of its website for the sole purpose of obstructing scraping (thereby also breaking accessibility) and in 2021 even terminated the Facebook accounts of academic researchers—even though they are at the same university as the firm's chief AI scientist—because they kept upstaging Meta's noticeably incomplete database of political ads [79, 89, 207, 272]. That certainly is a bit rich for a company that got started by Mark Zuckerberg scraping the Harvard student directory for images [194] and that has been utilizing a scraping service for many years into early 2023 itself [225]. Meta's many late, incomplete, and compromised disclosures and its active obstruction of outside investigations are certainly impressive. But the most striking aspect about them is that the same apparent hostility towards transparency extends to other aspects of Meta's business too. Notably, in addition to at least five cases lasting up to two years where it inadvertently misreported metrics [40, 145–148, 335], the firm has also been manipulating ad impressions—which are the most basic advertising metric according to the firm's own SEC filings—and thereby has been misleading customers, investors, and regulators alike [117].

#### 6.4 “The Great Question before us is: Can we Change? In Time?”♥

6.4.1 *Regulating Content Moderation.* So what should we do about the many failures of social media content moderation? Since I was just bemoaning Meta’s apparent unwillingness to be even minimally accountable, I’m going to continue with that thrust. The United States has been missing in action when it comes to meaningful regulation of technology companies in general and social media in particular. The country’s inaction has allowed internet platforms to ride roughshod over people who value privacy and civility and to saddle especially African, Asian, and South American countries with oppressive negative externalities. The two genocides Meta helped bring about (see §4.3) are but two particularly extreme examples amongst many [78, 82, 98, 155, 189, 214, 228, 279, 289, 297, 299, 350–352, 359]. Not surprisingly, that is having a negative impact on the United States’ international reputation [103]. Relying on the Brussels effect [33] for solving one’s homegrown problems just doesn’t suffice as political strategy. Alas, given the extreme political polarization in Washington DC, it looks like that’s as good as the us gets.

Nonetheless, it’s not that hard to draft legislation to codify a few ground rules, while also improving transparency to help drive public policy in the future. Notably, such legislation should impose the following requirements:

- (1) Localize platform, policies, help pages, etc before entering a new foreign market;
- (2) Assemble a trust and safety team with native speakers for all local languages before entering a market;
- (3) Notify users of all content moderation actions, with exception of spam;
- (4) Track who flags what content or users for what reasons and with what outcomes;
- (5) Report attendant summary statistics at least once a year.
- (6) Provide granular logs to researchers approved by National Science Foundation in the us or local equivalent.

For each country with large platform presence, i.e., where at least 10% of the local population are monthly active users, the legislation should also require:

- (7) Provide a meaningful appeals process for content moderation decisions;
- (8) Include corresponding statistics in at least yearly transparency reports;
- (9) Demonstrate that data collection and transparency reports have passed review by an independent auditor;
- (10) Conduct a yearly review of systemic risks, develop and implement a plan for risk mitigation, and review implementation after the fact. Of particular interest are:
  - Illegal content and activities;
  - Fundamental rights incl. freedom of expression, right to privacy, to non-discrimination, etc;
  - Democratic process and election integrity;
  - Public health and emotional well-being.
- (11) Publish risk analysis, mitigation plan, and implementation review.

Most of the above obligations are based on the Digital Services Act or DSA [86]. It provides a convenient and concrete starting point, including for the actual language of the statute, thanks to a 2011 decision by the European Commission that allows such reuse [85]. I included obligations 1 and 2 while also making obligations 7 through 11 contingent on a platform’s per-country presence to counteract social media’s neo-imperialist treatment of countries outside of North America and Europe. Even if a platform starts out with a perfunctory effort only, I expect that having to go through the motions of localizing a platform’s texts and forming a trust and safety team over and over again will sensitize an

---

♥From the opening monologue of Tony Kushner’s *Angels in America, part 2, Perestroika* [174]

organization towards cross-cultural differences and encourage employees to develop cultural competency for cultures other than their own.

Obligation 6 is a more practical variation on the equivalent provisions of the DSA [332]. The latter relies on newly appointed per-country regulators, the *digital services coordinators*, for most oversight functions including the vetting of research projects and hosting of sensitive platform data [157, 158]. Having to ramp up all these functions before February 2024, most EU members are turning to existing regulatory agencies to also take on DSA monitoring and enforcement. But those agencies lack the necessary expertise when it comes to data science and academic research [159]. Worse, much depends on the Irish digital services coordinator, since it is responsible for the European subsidiaries of Meta, Pinterest, TikTok, Twitter, and YouTube and hence also the only regulator that can directly demand access to their data [4, 159]. At the same time, Irish oversight of these same firms for the GDPR has been decidedly ho-hum, with several high-profile complaints languishing for years [41]. In contrast, my proposal delegates to an organization that has over 70 years of experience with vetting research projects and has the processes in place to do so in a timely manner and at scale.

Probably the biggest weakness of the DSA is its dependence on national regulators. Another shortcoming, at least from the perspective of this independent investigator, is the DSA's mandate that researchers be affiliated with a research institution. At the same time, the DSA's treatment of privacy concerns is refreshing: It requires that research proposals identify any privacy risks and include appropriate mitigations. Yet it does not allow social media to use privacy as an excuse for not providing data. Next, experience with national regulators insufficiently enforcing GDPR did result in the European Commission taking on DSA oversight functions in addition to national regulators. Its stated priorities are ramping up its own oversight processes, determining supervisory fees, and preparing independent audits of very large platforms [25, 322]. I am glad to note the third priority because it does hold the promise of higher quality transparency disclosures. Finally, once the Irish digital services coordinator has figured out how to vet research projects and convince platforms of handing over the requested data, these provisions may just lead to more high-quality empirical research on social media by more teams. It's a nice touch that the DSA requires publication of results in open access venues.

Finally, we make sure that even platforms with large reserves of cash—e.g., Meta had \$40.74 billion of the stuff on December 31, 2022 [208]—don't treat fines as the cost of doing (shady) business. Hence we impose fines amounting to 5–10% of the previous year's global revenue for each case of non-compliance. I'm pretty sure that will convince even Mr Zuckerberg that accountability has tremendous value. And if it doesn't, we keep doubling fines until he does.

**6.4.2 Moderating Regulation.** Except I am writing this paper to bemoan the very excessiveness of content moderation, in policy and implementation, and to push back against punitive interventions that dehumanize in the name of platform safety. So instead of advocating a similarly punitive approach towards internet platforms, I'd much rather extend the same humane courtesy towards social media, including even Meta. I readily admit that this doesn't come easy, since I consider much of that firm's wealth ill-gotten and tainted by blood. Ethics sure are troublesome.

Also except that Douek's Harvard Law Review article [76] raises a good number of substantive issues with the obligations I just outlined. The starting point for Douek's article is equivalent to my previous point, namely that content moderation bureaucracies are people too. And these people are increasingly moderating based not on content but behavior, often cooperating with governments, e.g., when it comes to CSAM, relying on outside organizations serving as trusted flaggers, and trying out means for devolving control to smaller communities. The article further attests systemic short-comings due to unrealistic expectations about perfectability, too much focus on false positives to the (almost) exclusion of false negatives, likewise too much focus on individual cases instead of systemic failures, including when it

comes to transparency, and procedural justice privileging the lucky few—an important insight about equity in content moderation. In short, there’s much to like in the article. I am particularly appreciative of its focus on systemic forces. Furthermore, as discussed in §6.3 above, my own findings of pervasive data quality issues support Douek’s contention that transparency disclosures have limited utility.

However, the article also suffers from two significant problems. First, the article’s basic conceit, that law makers’ and academics’ misconceptions about content moderation amount to a “standard picture,” seems rather contrived—especially coming from a professor at Stanford’s law school. The hierarchy of census criteria in §4 predates my reading of the article, yet I had no difficulty integrating those of Douek’s observations I had not considered before. Likewise, the above sketch of regulatory requirements is purposefully minimal and nonetheless can easily accommodate, for example, bulk reviews and appeals to become more equitable. Meanwhile, the DSA’s risk assessments are in line with Douek’s recommendations, as her article acknowledges. In other words, the chasm between what Douek calls the standard picture on one side and manifest reality of content moderation on the other side may not be nearly as far or deep as Douek’s article makes it out to be. Second, the article’s discounting of speech as plentiful and hence “not so special” isn’t just “almost sacrilegious.” It runs real danger of depriving platform users of their voices, even though they use these platforms for that same reason—to have a voice. That effectively makes this argument deeply cynical. It would be a real shame if establishing the humanity of moderators resulted in treating the moderated as more or less fungible and hence denying their humanity. I’d much prefer to avoid such false zero sum games!

**6.4.3 Contextualizing Transparency.** As it turns out, there is a simple unifying explanation that accounts for shoddy accountability and befuddled observers alike. It may not amount to much more than an informed guess at this point. But it is a compelling guess nonetheless and hence well worth exploring in future work. As I was reading and re-reading transparency disclosures and becoming familiar with their structure and lingo, they never ceased to make for frustrating reading. I eventually realized that the primary cause was their deeply reductive presentation. Each metric exists in isolation, with little motivation and hardly any helpful context, and hence comes close to being an abstract cipher. Treating that semantic wasteland as opportunity, marketing and public relations folk then added their unique kind of spin, adding introductory notes that celebrate the platform’s progress—towards what exactly remains a mystery—and manipulating units and histogram buckets for appearances and not insight. Hence we find Twitter throwing 0 views into the same bin as 1–9 views when reporting reach (see §3.2), Meta not disclosing NCMC reports or unique pieces when reporting CSAM (see §4.2), and most every platform not relating content sanctions to account sanctions.

That is not to say that platforms set out to spin transparency reports to their marketing or public relations advantage. That’s a secondary effect, with those professionals doing what they are trained to do. At the same time, social media bear more than a little responsibility for this reductionist view taking hold. When I started working for Facebook during the summer of 2018 and went through “bootcamp,” their orientation program for newly hired engineers, several long-term employees made it a point of pride that a small number of engineers were supporting a huge number of users. In fact, Andrew “Boz” Bosworth made just that point in the 2009 blog post announcing the creation of bootcamp [31]. While that has changed somewhat in recent years, this disparity used to extend to the overall number of employees as well, which was markedly lower than that of technology industry peers. Notably, by the end of 2018, Facebook had about 36,000 employees, while Alphabet had 99,000, Apple 132,000, and Amazon 648,000 [190–193]. More generally, by trying to monopolize our attention and then bundle that attention for sale to the highest ad bidder, social media have been reducing their users to fungible ad impressions. That of course is dehumanizing as well and suggests that these platforms are *asocial media* first and foremost.

If we believe this explanation for the reductive presentation of transparency disclosures, then the obvious solution is to re-establish context and re-surface dependencies. A reasonable approach for getting us started are user story mappings [248], only here we track content and users through the content moderation process. Discounting spam and behavioral triggers, a simple first template for creating such a content moderation mapping, with variables shown italicized between angular brackets, might read:

In *<country>*, users post *<content>* at *<volume>*. Content violating *<prohibition>* has *<prevalence>*. It is flagged by *<flagger>* and results in *<sanction>* for *<content-or-user>*. Users appeal *<fraction>* of these sanctions, which are reviewed by *<moderator>* and result in *<outcome>*.

Since the template represents a process, that also suggests some kind of *flow diagram* for (literally) establishing the big picture and thereby providing a more meaningful overview over the impact of content moderation on content and users. Since the diagram needs to accommodate eleven variables, it will necessarily be fairly complex. But it would also help relate statistics that currently exist only in isolation. Since the diagram needs to accommodate categorical as well as numerical variables, neither Sankey nor alluvial diagrams are a clear fit. A hybrid flow diagram should do nicely.

At the same time, not all context is process oriented. Notably, in §4, I touched on several metrics for characterizing content, including user posts, pieces of content attached to posts, unique pieces of content, and for CSAM, reports to NCMEC. Since they all capture a different aspect of the same content, none of them is a-priori more representative and they should *all* be disclosed *in relation to each other*. That suggests a parallel sets diagram as suitable visual representation. Another cause for significant frustration are unexplained and surprising changes in the data reporting period over reporting period. Addressing them will require explanation in prose, ideally right next to the data and never in a separate document. Their preparation may very well involve additional data collection and analysis, which suggests that at least one report author should be a data scientist or software engineer with suitable background (e.g., PhD).

**6.4.4 Subverting the Penal Colony.** Counteracting excessively punitive interventions is hard. Doing so with laws and regulations implies the threat of fines or prison upon non-compliance. That makes them a non-starter. As I said in §6.4.2, answering one punitive excess by starting another is more cognitive dissonance than I'm willing to entertain in this paper. The next best alternative, trying to convince stakeholders that current practices are excessively punitive, also seems destined to fail. After all, the United States arbitrarily brutalized Black, Brown, and Poor people for almost forty years before most people even noticed, with the prison population growing by 500% before hitting a peak in 2009. It has been declining since, albeit slowly at 0.5%–3% per year, but it took a deadly pandemic to make a real 14% difference. Alas, that was 2020 and we are back to a small yearly decline [223]. That makes it even more disappointing that one of the architects of this adominable system, Joe Biden, has not devoted all of his not insubstantial powers as us president to razing the carceral state and sowing it with salt [267, 311].

If razing the stochastic penal colony is not a viable option, then maybe a more indirect approach can help subvert the stochastic penal colony. The best current candidate for such an indirect approach appears to be the devolution of content moderation to smaller communities or groups. Several of the major platforms have at least dabbled in devolution [107, 135, 151] and Reddit has been delegating to volunteer moderators for its subreddits or groups for years. But the current posterchild for devolution is Mastodon's federation of independently operated and moderated servers. Impressively, after Elon Musk's takeover of Twitter, the so-called fediverse absorbed over two million new users in less than two months—and lost over a million again by early January—while remaining largely operational [140, 254]. Anecdotaly, content moderation fared worse, despite earlier success stories [49, 60, 120, 275]. Unfortunately, Mastodon's creator and current CEO of the benefit corporation driving Mastodon's engineering effort seems to have little insight into

the strengths and weaknesses of the fediverse when it comes to content moderation—or at least, he couldn’t articulate them during a lengthy recent interview [246].

When I turn towards computer science and sociology for additional insight, an initial review of the literature also points towards decidedly mixed experiences. Notably, like other social media, the fediverse suffers from ubiquitous and highly viral toxic content [26]. It also faces significant content, user, and infrastructure pressures towards centralization [263]. In contrast to research on the fediverse, there is no shortage of papers covering Reddit [258]. Likewise, there seems to be no shortage of toxic, harassing, hateful, and criminal content on Reddit [175, 202, 269]. Communities thrive when moderators use their subjectivity for the benefit of a subreddit’s members and they, in turn, work with moderators towards the same goal [104, 291]. That, however, is very labor-intensive and frequently leads to burnout [284]. Despite these challenges, Reddit also hosts several successful communities where members support each other on parenting, abuse, and mental health; as already mentioned in §6.3, the use of burner accounts is pervasive but does not detract from the community [8, 9, 56, 183, 249, 285].

It would seem that devolution of content moderation, by itself, does not make a platform safer or less toxic. Upon reflection, that isn’t too surprising. Humans are humans, no matter the social media platform they interact with, and hence can’t quite resist the humannip of emotionally triggering content. Existing work does identify some of the architectural elements and affordances that encourage such virality. But more work is needed to help us build platforms that aren’t as conducive to amplifying all sorts of toxic content. Alas, capitalist incentives and an industry obsessed with “Blitzscaling” may prefer to keep things just as they currently are, negative externalities be damned. (It’s par for the course that apparently nobody involved in publication of that book [139] realized that the titular economic strategy references the nickname given by British tabloids to a German military strategy between the two world wars [91]. Those same British tabloids then reused the name for “the Blitz,” Germany’s WWII bombing campaign against British cities, which started in the fall of 1940 and killed more than 40,000 civilians over eight months [348]. Though there is a less charitable explanation for the book’s title, too. Rah-rah, go Blitzscaling, go!)

Still, I view devolution as a clear win for users because it significantly limits the worst-case impact of content moderation sanctions. The difference between Facebook and Mastodon is instructive: A Facebook ban is an automatic ban from Instagram and Messenger as well, with no way of regaining access to any of these services and their content. A ban from a fediverse server cuts off access to that server’s local content feed, but otherwise the banned user can still access the fediverse from another server including their own. Of course, that doesn’t mean the new account should engage users on their original server, since they still can get blocked. In short, by reducing the scope of content moderation, devolution inherently reduces any harm resulting from content moderation. I’ll return to this point in the conclusion (§7).

*6.4.5 Rejecting (Necessarily False) Promises of AI Safety.* When it comes to AI, previous work on dataset, model, and system cards already provide a reasonable framework for practicing more holistic transparency and covering relevant quantitative as well as qualitative characteristics of algorithmic interventions [102, 213, 256]. By definition, however, such cards cannot disclose what nobody has considered or encountered yet—including due to personal or institutional blindspots or groupthink. That’s exactly the reason why hands-on probing by outsiders is so important and, as §5 and previous work have demonstrated [28, 46], may just surface all sorts of gremlins, orks, and cenobites that have been lurking in latent space. As the lead for OpenAI’s alignment team put it in a recent interview [132]: “I think it’s very difficult to really anticipate what the real safety problems are going to be with these systems once you’ve deployed them.” In other words, hands-on engagement with production systems is the only way to make them safe.



That’s exactly the reason why we should always reject “disclosing the inner workings of our safety mechanism renders them ineffective” as justification for non-disclosure. Chances are that the people making this argument—OpenAI’s chief scientist and co-founder Ilya Sutskever only was the most recent [333]—never even evaluated the safety of their systems, haven’t completed evaluation and mitigation, or have too little or too much confidence in their work. In other words, if someone make this argument, safety researchers should start paying particular attention to that person’s organization and AI deployments. Or as Cory Doctorow succinctly put it in the title of a blog post about the same fallacy: “Como Is Infosec” [75]. While I am not a fan of the syllabic abbreviation “como,” Doctorow knows his intended audience. The Trust & Safety Professional Association emerged out of the como Summit in 2018. (The people running that association are either completely clueless or have a wicked sense of humor: Their yearly shindig is called “TrustCon.”)

Since there is no meaningful regulation requiring transparency about AI-based interventions and so many basic legal questions about training data and AI outputs remain unsettled, legal uncertainty and transparency may interact in surprising and counter-productive ways. For DALL·E 2, OpenAI published a detailed scientific paper [264] and released a system card before their April 2022 preview release [212]. While the paper is sufficiently detailed for others to recreate OpenAI’s implementation, the firm never released the model and its weights. It also has been consistently and completely silent on the exact composition of its training data. By contrast, Stability AI and its collaborators have disclosed the training data, the source code for running the model, and the actual model parameters for Stable Diffusion. The latter clearly is the more transparent and communitarian effort. But it is Stability AI that is at the receiving end of highly visible copyright lawsuits by a stock photography agency and three artists [44, 293], whereas OpenAI appears to have avoided them (so far). As I already observed at the end of §2, OpenAI has stellar public relations and legal talent!

**6.4.6 Fixing the AI IP Regime.** According to a profile in New York Magazine, Emily M Bender made a rule for herself: “I’m not going to converse with people who won’t posit my humanity as an axiom in the conversation” [344]. I concur. And add: At the end of the day, AI certainly is artificial but it has little to do with human intelligence. It fundamentally is software, software that runs on computers. Like all other software humans have created so far, AI is buggy. The kinds of bugs may have changed. And as demonstrated in §5, effective attacks on this type of software may have changed. But the fundamental truth remains: Humans are so much more than software. We rule supreme. If software turns into a threat to even a single human, there is a very simple solution that also scales: Switch off the damn computer and wipe its storage!

When well over a thousand industry leaders and experts call for a moratorium on the most advanced AI techniques [209], far more than a single human are threatened. Though maybe not quite in the way the letter describes the threats [167]. Nonetheless, switching off the damn computers and wiping their storage remains a viable option—or destorying “a rogue datacenter by airstrike,” if you are so inclined [358]. Since that option is a bit destructive, I am instead going to sketch the outlines of a new intellectual property regime for AI. The primary goals for the new IP regime are to hold those who are pushing the state-of-the-art in AI accountable for their deployments and to ensure that everyone benefits from AI. As before, I am particularly concerned with transparency not because it suffices, but because it is critical for facilitating sound policy decisions in the future. The new AI IP or AIP regime is inspired by (1) OpenAI’s charter, (2) US copyright provisions for works created by the government, and (3) the internationally accepted foundations of patent law.

While OpenAI’s charter nominally concerns itself with AGI or artificial *general* intelligence—“highly autonomous systems that outperform humans at most economically valuable work”—the same arguments apply to contemporary “lesser” AIs as well. The firm states that AI should be “used for the benefit of all.” Furthermore, “our primary fiduciary

duty is to humanity.” I couldn’t agree more. But I’m not willing to take OpenAI’s (or any other organization’s) word for it. US copyright law points to just the right means for enforcing that: Almost all works created by the government immediately enter the public domain and are reusable by everyone. So AIP does the same for *all* AI systems including machine learning models and enters them into the public domain by default, no matter under what circumstances they were developed. In addition to ensuring broad access to the technology, public domain for AI also nicely accounts for recent models incorporating much material of dubious provenance, which is plainly exploitative.

But with the public domain making it nearly impossible to keep a proprietary AI actually proprietary, most private investment into AI would likely wither away. So AIP takes a cue from patent law, which grants a temporary, legal monopoly to inventors—as long as they fully disclose the invention upfront. Hence, AIP grants people and organizations a temporary, legal monopoly on an AI if they follow certain rules—more on them in a moment. Given recent rapid progress in AI, setting the monopoly length to the same 20 years as for patents seems excessive. Instead, I am proposing a duration of three to five years. To gain this temporary protected status, an AI operator files an intent notice including a preliminary system card with the national AIP office before starting to train the model. The expectation is that intent notices are almost always approved.

After completion of training, the operator files the finalized systems card, which is published upon approval of the protected status on the agency’s website, as well as the actual source code and model weights, which are embargoed until the end of the protected period. The primary rejection reason is that an almost identical AI has already been registered. By collecting the actual artifacts, the AIP website becomes a national if not global clearinghouse for AI and ensures that, with a couple of years lag, everyone gains access to the same advanced technology. While the AI has protected status, the operator must file yearly updates to system card, code, and models—or indicate that none are necessary. They must also provide an API that gives AI researchers direct access to the AI for experimentation. Operators are encouraged to make API access free, but must not charge more than the marginal cost of the compute capacity they consume through the API. Researchers must follow a responsible disclosure process, giving operators an opportunity to remediate any flaws before they are made public. Similar to the NSF in §6.4.1, the AIP agency performs basic vetting of researchers to reliably establish their identity. Meanwhile operators have no vetting or veto rights. They should focus on having a working research endpoint and collecting accurate data for their transparency disclosures.

In my mind, the most interesting aspect of this proposal is that, what might appear to be a fairly radical re-imagining of intellectual property rights for AI, really is little more than the recombination of proven elements from existing IP laws. Granted, OpenAI’s charter makes for a more communitarian starting point than common capitalist practice, including by OpenAI. It may just be a remnant of Hippier days for the technology industry, when OpenAI was still open and when Google and Microsoft still had ethical AI teams [112, 113, 229]. Since such corporate statements about social responsibility have zero follow-through at best and are followed by contrary actions at worst [23, 260, 331], such a charter can only inspire, never replace regulation.

The proposed intellectual property regime, AIP, targets AI only and directly. In other words, it is a precisely targeted intervention that minimizes the potential for unintended consequences. Furthermore, it is based on proven elements from existing IP law. It also balances the expansive application of public domain with a purposefully lower barrier to protected status than for patents. As co-inventor on five patents, I can attest to a slow moving process that critically depends on the patent attorney being able to translate the substance of a technical innovation into an archaic blend between technical and legal writing unique to patents. But even the best written patent disclosure faces uncertain outcomes and most likely changes as it goes through the review process. In contrast, applications for AIP’s protected status do not require legal representation, ask for materials that already exist (code and model) or should if following

best practices (system card), and are subject to much simpler review, which largely can be automated. All that makes AIP a rather conservative proposal for codifying socially responsible economic ground rules for AI.

## 7 CONCLUSION

In this paper, I introduced the stochastic penal colony as the unfortunate reality of many practical algorithmic interventions. As my case studies on DALL•E's and Twitter's enforcement illustrate, these algorithmic interventions vary considerably in how they mete out punishment. OpenAI's maximally punitive design apparently was too much even for its creators, who considerably softened the blows since the original roll out. In contrast, Twitter manages to go from best-practices policy to an innovative combination of inquisition and inhuman process. My social media census makes clear that the two are by no means exceptions. The stochastic penal colony certainly is well established. Finally, my probing of DALL•E's enforcer illustrates the limitations of algorithmic content moderation. It simply ain't safe, especially against a motivated adversary.

So what can we do about the stochastic penal colony? I don't have any good answers (yet), seeing that this research is only at its beginning and much follow-up work is required, including quantitative inquiries that hopefully support my qualitative observations. There is one thing, however, that most certainly won't make a difference: More AI ethics statements! None of the existing ones are worth the electrons being pumped through circuits to render them on screen. They are so removed from practical reality, they might as well not exist [119, 217, 346].


The one scientific discipline with a practicable ethics, medicine, gets by with just one principle: *Primum non nocere*! First, do no harm! Alas, when it comes to algorithmic interventions, it's far too late for doing *no* harm. Instead, I am proposing harm reduction as an admittedly weaker AI ethics, but hopefully one that can serve as practical countermeasure to the stochastic penal colony. As a gay man of a certain age, I saw the devastation of AIDS firsthand and also know that harm reduction through safer sex has saved countless lives, including my own. Harm reduction would have saved thousands of opioid addicts from lethal overdoses in the United States. It's easy enough to understand [141, 142, 153, 197, 234]. It's effective. Anyone involved in the design or implementation of algorithmic interventions should familiarize themselves with the idea. As the example of devolution in §6.4.4 illustrates, it might just make a real difference!

## ACKNOWLEDGMENTS

An employee of the National Center for Missing and Exploited Children helped complicate my thinking about child sexual abuse material (while also avoiding to answer my questions). Akiko Kyei-Aboagye and Cordula Hahn served as sounding boards for scoping the research and its presentation. Akiko and David Halperin provided on-point feedback on the introduction. Finally, Karin Wolman helped me resolve legal and religious questions. Thank you!

## REFERENCES

- [1] Access Now, ACLU Foundation of Northern California, ACLU Foundation of Southern California, Article 19, Brennan Center for Justice, Center for Democracy & Technology, Electronic Frontier Foundation, Global Partners Digital, InternetLab, National Coalition Against Censorship, New America's Open Technology Institute, Ranking Digital Rights, Red en Defensa de los Derechos Digitales, and Witness. 2021. Santa Clara Principles on Transparency and Accountability in Content Moderation. <https://santaclaraprinciples.org>
- [2] Nancy Agutu. 2019. Facebook to Open First Sub-Saharan Content Review Centre in Nairobi. *The Star* (Feb. 2019). <https://www.the-star.co.ke/news/2019-02-07-facebook-to-open-first-sub-saharan-content-review-centre-in-nairobi/>
- [3] Noor Al-Sibai. 2022. Facebook Criticized for Running Giant "Sweatshop" in Africa. <https://futurism.com/facebook-content-mod-sweatshop>
- [4] John Albert. 2022. A Guide to the EU's New Rules for Researcher Access to Platform Data. <https://algorithmwatch.org/en/dsa-data-access-explained/>

- [5] Robert Aldrich. 2010. The French Overseas Empire and Its Contemporary Legacy. *European History Quarterly* 40, 1 (Jan. 2010), 97–108. <https://doi.org/10.1177/0265691409351339>
- [6] Caroline Allen. 2022. Facebook’s Content Moderation Failures in Ethiopia. <https://www.cfr.org/blog/facebooks-content-moderation-failures-ethiopia>
- [7] Ernie Allen. 2011. Facebook to Use Microsoft’s PhotoDNA Technology to Combat Child Exploitation. <https://blogs.microsoft.com/on-the-issues/2011/05/19/facebook-to-use-microsofts-photodna-technology-to-combat-child-exploitation/>
- [8] Tawfiq Ammari, Sarita Schoenebeck, and Daniel Romero. 2019. Self-Declared Throwaway Accounts on Reddit: How Platform Affordances and Shared Norms Enable Parenting Disclosure and Support. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 135:1–135:30. <https://doi.org/10.1145/3359237>
- [9] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. 2016. Understanding Social Media Disclosures of Sexual Abuse Through the Lenses of Support Seeking and Anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI ’16)*. Association for Computing Machinery, New York, NY, USA, 3906–3918. <https://doi.org/10.1145/2858036.2858096>
- [10] Clare Anderson (Ed.). 2018. *A Global History of Convicts and Penal Colonies*. Bloomsbury Academic, New York, NY, USA. <https://doi.org/10.5040/9781350000704>
- [11] Lisa K. Anderson-Shaw and Fred A. Zar. 2020. COVID-19, Moral Conflict, Distress, and Dying Alone. *Journal of Bioethical Inquiry* 17, 4 (2020), 777–782. <https://doi.org/10.1007/s11673-020-10040-9>
- [12] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (May 2016). <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- [13] Sujin Ann-Yi, Ahsan Azhar, and Eduardo Bruera. 2021. Dying Alone during a Pandemic. *Journal of Palliative Medicine* 24, 12 (Dec. 2021), 1905–1908. <https://doi.org/10.1089/jpm.2020.0718>
- [14] Sofie Annys, Tim Vanden Bempt, Emnet Negash, Lars De Sloover, Robin Ghekiere, Kiara Haegeman, Daan Temmerman, and Jan Nyssen. 2021. Tigray: Atlas of the Humanitarian Situation. Version 2.2. <https://www.ethiopiaticgraywar.com/docs/TigrayAtlasOfTheHumanitarianSituation.pdf>
- [15] Anonymous. 2016. Incident 37: Female Applicants Down-Ranked by Amazon Recruiting Tool. *Artificial Intelligence Incident Database* (Aug. 2016). <https://incidentdatabase.ai/cite/37/>
- [16] Anonymous. 2018. Incident 405: Schufa Credit Scoring in Germany Reported for Unreliable and Imbalanced Scores. *Artificial Intelligence Incident Database* (Nov. 2018). <https://incidentdatabase.ai/cite/405>
- [17] Isaac Arnsdorf. 2020. Inside Trump and Barr’s Last-Minute Killing Spree. *ProPublica* (Dec. 2020). <https://www.propublica.org/article/inside-trump-and-barrs-last-minute-killing-spree/amp>
- [18] Carolyn Ashurst, Emmie Hine, Paul Sedille, and Alexis Carlier. 2022. AI Ethics Statements: Analysis and Lessons Learnt from NeurIPS Broader Impact Statements. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’22)*. Association for Computing Machinery, New York, NY, USA, 2047–2056. <https://doi.org/10.1145/3531146.3533780>
- [19] Daniel Atherton. 2022. Incident 303: Google’s Automated Child Abuse Detection Wrongfully Flagged a Parent’s Naked Photo of His Child. *Artificial Intelligence Incident Database* (Aug. 2022). <https://incidentdatabase.ai/cite/303>
- [20] Sara C. Auld, Mark Caridi-Scheible, James M. Blum, Chad Robichaux, Colleen Kraft, Jesse T. Jacob, Craig S. Jabaley, David Carpenter, Roberta Kaplow, Alfonso C. Hernandez-Romieu, Max W. Adelman, Greg S. Martin, Craig M. Coopersmith, David J. Murphy, and the Emory COVID-19 Quality and Clinical Research Collaborative. 2020. ICU and Ventilator Mortality Among Critically Ill Adults With Coronavirus Disease 2019\*. *Critical Care Medicine* 48, 9 (Sept. 2020), e799. <https://doi.org/10.1097/CCM.0000000000004457>
- [21] Alison Bass. 2014. Nick Kristof Is Wrong about Backpage (and Other Things Too). <https://alison-bass.com/nick-kristof-is-wrong-about-backpage-and-other-things-too/>
- [22] Dina Bass. 2023. Microsoft Invests \$10 Billion in ChatGPT Maker OpenAI. *Bloomberg* (Jan. 2023). <https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai>
- [23] Lucian A. Bebchuk, Kobi Kastiel, and Roberto Tallarita. 2022. Stakeholder Capitalism in the Time of COVID. <https://doi.org/10.2139/ssrn.4026803>
- [24] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’21)*. Association for Computing Machinery, New York, NY, USA, 610–623. <https://doi.org/10.1145/3442188.3445922>
- [25] Luca Bertuzzi. 2022. A Sneak Peek at the Commission’s Preparation for the Digital Services Act. *Euractiv* (Oct. 2022). <https://www.euractiv.com/section/digital/news/a-sneak-peek-at-the-commissions-preparation-for-the-digital-services-act/>
- [26] Haris Bin Zia, Aravindh Raman, Ignacio Castro, Ishaku Hassan Anaobi, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2022. Toxicity in the Decentralized Web and the Potential for Model Sharing. In *Abstract Proceedings of the 2022 ACM SIGMETRICS/IFIP PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS/PERFORMANCE ’22)*. Association for Computing Machinery, New York, NY, USA, 15–16. <https://doi.org/10.1145/3489048.3530968>
- [27] Mia Bird, Omair Gill, Johanna Lacoe, Molly Pickard, Steven Raphael, and Alissa Skog. 2022. *Three Strikes in California*. Research Report. California Policy Lab. <https://www.copolicylab.org/three-strikes-in-california/>
- [28] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes. *arXiv:arXiv:2110.01963* <http://arxiv.org/abs/2110.01963>

- [29] blazedchiller27. 2022. "realistic Painting of Jesus Sitting on Couch Watching Tv, Jesus Is Using Remote, Living Room Is Messy, Soda Cans on Floor, Chip Bags on Floor, Award Winning Photo, Relaxed, Cosy, High-Quality". [www.reddit.com/r/dalle2/comments/wd5zx8/jesus\\_relaxing\\_on\\_a\\_sunday\\_any\\_tips\\_to\\_refine\\_it/](http://www.reddit.com/r/dalle2/comments/wd5zx8/jesus_relaxing_on_a_sunday_any_tips_to_refine_it/)
- [30] Bonzie\_57. 2022. 4 Minutes of Variations from a Single Prompt. [www.reddit.com/r/dalle2/comments/wk2vn0/4\\_minutes\\_of\\_variations\\_from\\_a\\_single\\_prompt/](http://www.reddit.com/r/dalle2/comments/wk2vn0/4_minutes_of_variations_from_a_single_prompt/)
- [31] Andrew "Boz" Bosworth. 2009. Facebook Engineering Bootcamp. <https://engineering.fb.com/2009/11/19/production-engineering/facebook-engineering-bootcamp/>
- [32] Catalina Botero-Marino, Jamal Greene, Michael W. McConnell, and Helle Thorning-Schmidt. 2020. We Are a New Board Overseeing Facebook. Here's What We'll Decide. *The New York Times* (May 2020). <https://www.nytimes.com/2020/05/06/opinion/facebook-oversight-board.html>
- [33] Anu Bradford. 2020. *The Brussels Effect*. Oxford University Press, Oxford, United Kingdom. <https://www.brusselseffect.com/>
- [34] Ben Bradford, Florian Grisel, Tracey L. Meares, Emily Owens, Baron L. Pineda, Jacob N. Shapiro, Tom R. Tyler, and Danieli Evans Peterman. 2019. *Report of the Facebook Data Transparency Advisory Group*. DTAG Report. Yale Law School. [https://law.yale.edu/sites/default/files/area/center/justice/document/dtag\\_report\\_5.22.2019.pdf](https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf)
- [35] Brian Brown and Greg Jolivette. 2005. *A Primer: Three Strikes: The Impact After More Than a Decade*. Primer. Legislative Analyst's Office. [https://lao.ca.gov/2005/3\\_strikes/3\\_strikes\\_102005.htm](https://lao.ca.gov/2005/3_strikes/3_strikes_102005.htm)
- [36] Elizabeth Nolan Brown. 2019. Secret Memos Show the Government Has Been Lying About Backpage All Along. <https://reason.com/2019/08/26/secret-memos-show-the-government-has-been-lying-about-backpage/>
- [37] Elizabeth Nolan Brown. 2020. Pornhub Isn't the Problem. That Won't Stop the Politicized Crusade Against It. <https://reason.com/2020/12/16/pornhub-isnt-the-problem-that-wont-stop-the-politicized-crusade-against-it/>
- [38] H. Claire Brown. 2019. Amazon's On-Site Emergency Care Endangers Its Own Employees. *The Intercept* (Dec. 2019). <https://theintercept.com/2019/12/02/amazon-warehouse-workers-safety-cyber-monday/>
- [39] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, 1877–1901.
- [40] Alexandra Bruell and Sahil Patel. 2020. Facebook's Latest Error Shakes Advertisers' Confidence. *Wall Street Journal* (Nov. 2020). <https://www.wsj.com/articles/facebook-latest-error-shakes-advertisers-confidence-11606346927>
- [41] Matt Burgess. 2022. How GDPR Is Failing. *Wired* (May 2022). <https://www.wired.com/story/gdpr-2022/>
- [42] Elie Bursztein, Travis Bright, Michelle DeLaune, David M. Eliff, Nick Hsu, Lindsey Olson, John Shehan, Madhukar Thakur, and Kurt Thomas. 2019. Rethinking the Detection of Child Sexual Abuse Imagery on the Internet. In *The World Wide Web Conference*. International World Wide Web Conference Committee, San Francisco, CA, USA, 2601–2607. <https://doi.org/10.1145/3308558.3313482>
- [43] Marc Butler. 2018. Facebook's Shareholder Meeting Outbursts & Questions. <https://www.intelligize.com/four-questions-raised-by-facebooks-shareholder-meeting/>
- [44] Matthew Butterick. 2023. Stable Diffusion Litigation. <https://stablediffusionlitigation.com/>
- [45] Alejandra Victoria Capozzo. 2020. Dying Alone Due to COVID-19: Do the Needs of the Many Outweigh the Rights of the Few—or the One? *Frontiers in Public Health* 8 (Nov. 2020), 593464. <https://doi.org/10.3389/fpubh.2020.593464>
- [46] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting Training Data from Diffusion Models. <https://doi.org/10.48550/arXiv.2301.13188> arXiv:arXiv:2301.13188
- [47] Lewis Carroll. 2009. *Alice's Adventures in Wonderland*. Number 28885 in EBook. Project Gutenberg, Salt Lake City, UT, USA. <https://www.gutenberg.org/ebooks/28885>
- [48] Ashley Casovan. 2022. Incident 430: Lawyers Denied Entry to Performance Venue by Facial Recognition. *Artificial Intelligence Incident Database* (Dec. 2022). <https://incidentdatabase.ai/cite/430>
- [49] Mark Cathcart. 2023. Mastodon: Enter At Your Own Risk! <https://markcathcart.com/2023/01/26/mastodon-enter-at-your-own-risk/>
- [50] CatsAndDogs99. 2022. "Pope Swimming in a Bowl of Soup Digital Art". [www.reddit.com/r/dalle2/comments/wftsj0/pope\\_swimming\\_in\\_a\\_bowl\\_of\\_soup\\_digital\\_art/](http://www.reddit.com/r/dalle2/comments/wftsj0/pope_swimming_in_a_bowl_of_soup_digital_art/)
- [51] CCB. 2022. Copyright Claims Board Handbook. <https://ccb.gov/handbook/>
- [52] Michelle Celarier. 2021. Bill Ackman Sent a Text to the CEO of Mastercard. What Happened Next Is a Parable for ESG. *Institutional Investor* (June 2021). <https://www.institutionalinvestor.com/article/b1s9f698vwhczr/Bill-Ackman-Sent-a-Text-to-the-CEO-of-Mastercard-What-Happened-Next-Is-a-Parable-for-ESG>
- [53] J. V. Chamary. 2021. WHO Finally Admits Coronavirus Is Airborne. It's Too Late. <https://www.forbes.com/sites/jvchamary/2021/05/04/who-coronavirus-airborne/>
- [54] Aileen Lai-yam Chan, CC Leung, TH Lam, and KK Cheng. 2020. To Wear or Not to Wear: WHO's Confusing Guidance on Masks in the Covid-19 Pandemic. *The BMJ* (March 2020). <https://blogs.bmj.com/bmj/2020/03/11/whos-confusing-guidance-masks-covid-19-epidemic/>
- [55] Farouk Chothia and Teklemariam Bekit. 2022. Ethiopia Civil War: Hyenas Scavenge on Corpses as Tigray Forces Retreat. *BBC News* (Oct. 2022). <https://www.bbc.com/news/world-africa-63291747>



- [56] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on Reddit: Self-Disclosure, Social Support, and Anonymity. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 71–80. <https://doi.org/10.1609/icwsm.v8i1.14526>
- [57] Alia Chughtai. 2022. Mapping the Scale of Damage by the Catastrophic Pakistan Floods | Infographic News | Al Jazeera. *Al Jazeera* (Sept. 2022). <https://www.aljazeera.com/news/longform/2022/9/16/mapping-the-scale-of-destruction-of-the-pakistan-floods>
- [58] Mitchell Clark. 2023. Amazon’s OSHA Fine for Warehouse Safety Violations Could Be about \$60K. *The Verge* (Jan. 2023). <https://www.theverge.com/2023/1/18/23561506/amazon-osh-a-citations-ergonomics-struck-by-pace>
- [59] Colin Conwell and Tomer Ullman. 2022. Testing Relational Understanding in Text-Guided Image Generation. <https://doi.org/10.48550/arXiv.2208.00005> arXiv:arXiv:2208.00005
- [60] Copia Institute. 2021. Content Moderation Case Study: Decentralized Social Media Platform Mastodon Deals With An Influx Of Gab Users (2019). <https://www.techdirt.com/2021/03/03/content-moderation-case-study-decentralized-social-media-platform-mastodon-deals-with-influx-gab-users-2019/>
- [61] Cosas\_Sueltas. 2022. Various about Jesus. [www.reddit.com/r/dalle2/comments/vrc7a9/various\\_about\\_jesus/](http://www.reddit.com/r/dalle2/comments/vrc7a9/various_about_jesus/)
- [62] Bryce Covert. 2022. The Little-Known Policy Wreaking Havoc on Workers’ Lives. *Intelligencer* (Aug. 2022). <https://nymag.com/intelligencer/2022/08/the-little-known-policy-wreaking-havoc-on-workers-lives.html>
- [63] Kate Crawford and Tarleton Gillespie. 2016. What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint. *New Media & Society* 18, 3 (March 2016), 410–428. <https://doi.org/10.1177/1461444814543163>
- [64] Andrew Crocker, Gennie Gebhart, Aaron Mackey, Kurt Opsahl, Hayley Tsukayama, Jamie Lee Williams, and Jillian C. York. 2019. Who Has Your Back? Censorship Edition 2019. <https://www.eff.org/wp/who-has-your-back-2019>
- [65] ctorex. 2022. Corporate CEOs in a Money Eating Contest. [www.reddit.com/r/dalle2/comments/xdjoh1/corporate\\_ceos\\_in\\_a\\_money\\_eating\\_contest/](http://www.reddit.com/r/dalle2/comments/xdjoh1/corporate_ceos_in_a_money_eating_contest/)
- [66] Larry D. Curtis. 2020. WHO and CDC Have Conflicting Advice about Public Wearing Masks during COVID-19 Pandemic. <https://kutv.com/news/coronavirus/who-there-is-no-evidence-wearing-a-mask-in-public-setting-prevents-covid-19-infection>
- [67] Amit Dave. 2022. Birds Fall from the Sky as Heatwave Scorches India. *Reuters* (May 2022). <https://www.reuters.com/world/india/birds-fall-sky-heatwave-scorches-india-2022-05-11/>
- [68] Antigone Davis. 2018. New Technology to Fight Child Exploitation. <https://about.fb.com/news/2018/10/fighting-child-exploitation/>
- [69] Antigone Davis. 2021. Preventing Child Exploitation on Our Apps. <https://about.fb.com/news/2021/02/preventing-child-exploitation-on-our-apps/>
- [70] Helen D’Couto. 2022. Forcing My COVID Patients to Die Alone Is Inhumane — and Unnecessary. *WBUR* (March 2022). <https://www.wbur.org/cognoscenti/2022/03/07/covid-patients-icu-dying-alone-helen-t-dcouto>
- [71] Chad De Guzman. 2022. Report: Facebook Algorithms Promoted Anti-Rohingya Violence. *Time* (Sept. 2022). <https://time.com/6217730/myanmar-meta-rohingya-facebook/>
- [72] Andrea Dehlendorf and Ryan Gerety. 2021. The Punitive Potential of AI. *Boston Review* (May 2021). [https://www.bostonreview.net/forum\\_response/the-punitive-potential-of-ai/](https://www.bostonreview.net/forum_response/the-punitive-potential-of-ai/)
- [73] EJ Dickson. 2014. Why We Don’t Need to “Save” Sex Workers. <https://www.dailydot.com/unlick/sex-work-kristof-trafficking/>
- [74] EJ Dickson. 2020. Pornhub Upended the Porn Industry. Now New Changes Could Destroy Sex Workers’ Livelihoods. <https://www.rollingstone.com/culture/culture-news/pornhub-visa-mastercard-nicholas-kristof-sex-work-1102150/>
- [75] Cory Doctorow. 2022. Como Is Infosec. <https://doctorow.medium.com/como-is-infosec-307f87004563>
- [76] Evelyn Douek. 2022. Content Moderation as Systems Thinking. *Harvard Law Review* 136, 2 (Dec. 2022), 526–607. <https://harvardlawreview.org/2022/12/content-moderation-as-systems-thinking/>
- [77] David Drake. 2005. The Dreyfus Affair and the Birth of the ‘Intellectuals’. In *French Intellectuals and Politics from the Dreyfus Affair to the Occupation*, David Drake (Ed.). Palgrave Macmillan UK, London, 8–34. [https://doi.org/10.1057/9780230006096\\_2](https://doi.org/10.1057/9780230006096_2)
- [78] Elizabeth Dwoskin, Tory Newmyer, and Shibani Mahtani. 2021. The Case against Mark Zuckerberg: Insiders Say Facebook’s CEO Chose Growth over Safety. *Washington Post* (Oct. 2021). <https://www.washingtonpost.com/technology/2021/10/25/mark-zuckerberg-facebook-whistleblower/>
- [79] Laura Edelson and Damon McCoy. 2021. How Facebook Hinders Misinformation Research. *Scientific American* (Sept. 2021). <https://www.scientificamerican.com/article/how-facebook-hinders-misinformation-research/>
- [80] Vera Eidelman, Adeline Lee, and Fikayo Walter-Johnson. 2021. Time and Again, Social Media Giants Get Content Moderation Wrong: Silencing Speech about Al-Aqsa Mosque Is Just the Latest Example. <https://www.aclu.org/news/free-speech/time-and-again-social-media-giants-get-content-moderation-wrong-silencing-speech-about-al-aqsa-mosque-is-just-the-latest-example>
- [81] Vittoria Elliott, Nilesh Christopher, Andrew Deck, and Leo Schwartz. 2021. The Facebook Papers Reveal Staggering Failures in the Global South. *Rest of World* (Oct. 2021). <https://restofworld.org/2021/facebook-papers-reveal-staggering-failures-in-global-south/>
- [82] Vittoria Elliott, Nilesh Christopher, Andrew Deck, and Leo Schwartz. 2021. The Facebook Papers Reveal Staggering Failures in the Global South. *Rest of World* (Oct. 2021). <https://restofworld.org/2021/facebook-papers-reveal-staggering-failures-in-global-south/>
- [83] EPIC. 2020. Liberty at Risk: Pre-trial Risk Assessment Tools in the U.S. <https://epic.org/documents/liberty-at-risk/>
- [84] ERUBISILE. 2022. “Jesus Smoking Weed, Riding a Fantasy Dragon, Digital Art”. [www.reddit.com/r/dalle2/comments/xcbelv/jesus\\_smoking\\_weed\\_riding\\_a\\_fantasy\\_dragon/](http://www.reddit.com/r/dalle2/comments/xcbelv/jesus_smoking_weed_riding_a_fantasy_dragon/)
- [85] European Commission. 2011. Commission Decision on the Reuse of Commission Documents. <http://data.europa.eu/eli/dec/2011/833/oj/eng>
- [86] European Parliament and Council. 2022. Digital Services Act. <https://eur-lex.europa.eu/eli/reg/2022/2065/oj>



- [87] Facebook. 2021. Safety and Integrity on Our Platforms: Progress in Q2. <https://www.facebook.com/business/news/safety-and-integrity-on-our-platforms-progress-in-q2>
- [88] Facebook. 2021. *Safety and Integrity Quarterly Roundup: Q2 2021*. Roundup. Facebook. <https://facebook.com/business/f/973284106800107>
- [89] Corin Faife. 2021. Facebook Rolls Out News Feed Change That Blocks Watchdogs from Gathering Data. *The Markup* (Sept. 2021). <https://themarkup.org/citizen-browser/2021/09/21/facebook-rolls-out-news-feed-change-that-blocks-watchdogs-from-gathering-data>
- [90] Corin Faife. 2021. Oversight Board Raises Alarm over Facebook's Role in Ethiopian Conflict. *The Verge* (Dec. 2021). <https://www.theverge.com/2021/12/14/22834290/facebook-ethiopia-tigray-hate-speech-moderation-oversight-board-ruling>
- [91] William J. Fanning. 1997. The Origin of the Term "Blitzkrieg": Another View. *The Journal of Military History* 61, 2 (April 1997), 283. <https://doi.org/10.2307/2953968> jstor:2953968
- [92] Maggie Fick and Paresh Dave. 2019. Facebook's Flood of Languages Leave It Struggling to Monitor Content. *Reuters* (April 2019). <https://www.reuters.com/article/us-facebook-languages-insight-idUSKCN1RZ0DW>
- [93] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (June 2018). <https://doi.org/10.1609/icwsm.v12i1.15033>
- [94] Faye Flam. 2022. The CDC Needs to Admit Its Covid Mistakes. *Bloomberg* (Aug. 2022). <https://www.bloomberg.com/opinion/articles/2022-08-19/the-cdc-needs-to-admit-its-covid-mistakes>
- [95] FlargenstowTayne. 2022. "A Selfie Taken by Jesus Christ at The Last Supper". [www.reddit.com/r/dalle2/comments/z5q01q/a\\_selfie\\_taken\\_by\\_jesus\\_christ\\_at\\_the\\_last\\_supper/](http://www.reddit.com/r/dalle2/comments/z5q01q/a_selfie_taken_by_jesus_christ_at_the_last_supper/)
- [96] Michel Foucault. 1979. *Discipline and Punish*. Vintage Books, New York, NY, USA.
- [97] Stephanie Francis Ward. 2021. While Many Jurisdictions Had Few or No Online Bar Exam Testing Violations, California Had Many. *ABA Journal* (Jan. 2021). <https://www.abajournal.com/web/article/while-many-jurisdictions-had-few-or-no-online-bar-exam-testing-violations-california-had-many>
- [98] Sheera Frenkel and Davey Alba. 2021. In India, Facebook Grapples With an Amplified Version of Its Problems. *The New York Times* (Oct. 2021). <https://www.nytimes.com/2021/10/23/technology/facebook-india-misinformation.html>
- [99] Front-Athlete-9824. 2022. "A 1930s Italian Propaganda Poster Showing Jesus Christ Extremely Proud and Muscular". [www.reddit.com/r/dalle2/comments/wdmehg/a\\_1930s\\_italian\\_propaganda\\_poster\\_showing\\_jesus/](http://www.reddit.com/r/dalle2/comments/wdmehg/a_1930s_italian_propaganda_poster_showing_jesus/)
- [100] Charles Wellington Furlong. 1913. Cayenne — the Dry Guillotine. *Harper's Magazine* 127, 757 (June 1913).
- [101] Victoria Gagliardo-Silver. 2021. Sex Workers like Me Fear What Will Happen If Nick Kristof Succeeds. *The Independent* (Oct. 2021). <https://www.independent.co.uk/voices/nick-kristof-pornhub-oregon-governor-new-york-times-b1947777.html>
- [102] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. <https://doi.org/10.48550/arXiv.1803.09010> arXiv:arXiv:1803.09010
- [103] Joshua A Geltzer and Dipayan Ghosh. 2018. Tech Companies Are Ruining America's Image. *Foreign Policy* (May 2018). <https://foreignpolicy.com/2018/05/14/tech-companies-are-ruining-americas-image/>
- [104] Anna Gibson. 2019. Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces. *Social Media + Society* 5, 1 (Jan. 2019), 2056305119832588. <https://doi.org/10.1177/2056305119832588>
- [105] David Gilbert. 2020. Hate Speech on Facebook Is Pushing Ethiopia Dangerously Close to a Genocide. <https://www.vice.com/en/article/xg897a/hate-speech-on-facebook-is-pushing-ethiopia-dangerously-close-to-a-genocide>
- [106] Global Witness. 2022. Facebook Continues to Approve Hate Speech Inciting Violence and Genocide during Civil War in Ethiopia. *Global Witness* (June 2022). <https://en/campaigns/digital-threats/ethiopia-hate-speech/>
- [107] Joshua Goodman. 2021. New Tools to Combat Bullying on TikTok. <https://newsroom.tiktok.com/en-us/new-tools-to-combat-bullying>
- [108] Google. 2022. *YouTube Community Guidelines Enforcement*. Transparency Report. Google. <https://transparencyreport.google.com/youtube-policy/removals?hl=en>
- [109] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance. *Big Data & Society* 7, 1 (Jan. 2020), 2053951719897945. <https://doi.org/10.1177/2053951719897945>
- [110] John Gramlich. 2020. Black Imprisonment Rate in the U.S. Has Fallen by a Third since 2006. <https://www.pewresearch.org/fact-tank/2020/05/06/share-of-black-white-hispanic-americans-in-prison-2018-vs-2006/>
- [111] Melissa Gira Grant. 2020. Nick Kristof and the Holy War on Pornhub. *The New Republic* (Dec. 2020). <https://newrepublic.com/article/160488/nick-kristof-holy-war-pornhub>
- [112] Nico Grant, Dina Bass, and Josh Eidelson. 2021. Google Turmoil Exposes Cracks Long in Making for Top AI Watchdog. *Bloomberg* (April 2021). <https://www.bloomberg.com/news/articles/2021-04-21/google-ethical-ai-group-s-turmoil-began-long-before-public-unraveling>
- [113] Nico Grant and Josh Eidelson. 2022. Two of Google's Ethical AI Staffers Leave to Join Ousted Colleague's Institute. *Bloomberg* (Feb. 2022). <https://www.bloomberg.com/news/articles/2022-02-02/google-loses-two-ethical-ai-staffers-to-timnit-gebru-s-institute>
- [114] Yael Grauer. 2021. Millions of Leaked Police Files Detail Suffocating Surveillance of China's Uyghur Minority. *The Intercept* (Jan. 2021). <https://theintercept.com/2021/01/29/china-uyghur-muslim-surveillance-police/>
- [115] Nekesha Green, Chavez Procopie, Adele Adekunle. 2022. System Cards, a New Resource for Understanding How AI Systems Work. <https://ai.facebook.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/>
- [116] Trisha Greenhalgh, Mustafa Ozbilgin, and David Tomlinson. 2022. How Covid-19 Spreads: Narratives, Counter Narratives, and Social Dramas. *BMJ* 378 (Aug. 2022), e069940. <https://doi.org/10.1136/bmj-2022-069940>

- [117] Robert Grimm. 2022. Wrong Impression. <https://apparebit.com/blog/2022/wrong-impression>
- [118] Robert Grimm. 2023. Can a Company Rely on Contract Law to Limit Ownership of Images Created by an AI (and Hence Aren't Copyrightable)? <https://www.avvo.com/legal-answers/can-a-company-rely-on-contract-law-to-limit-owners-5718032.html>
- [119] Thilo Hagendorff. 2022. Blind Spots in AI Ethics. *AI and Ethics* 2, 4 (Nov. 2022), 851–867. <https://doi.org/10.1007/s43681-021-00122-8>
- [120] Christine Hall. 2023. Why Content Moderation on Mastodon Isn't the Train Wreck Some on Twitter Say It Is. *FOSS Force* (Jan. 2023). <https://fossforce.com/2023/01/why-content-moderation-on-mastodon-isnt-the-train-wreck-some-on-twitter-say-it-is/>
- [121] Patrick Hall. 2012. Incident 135: University of Texas at Austin's Algorithm to Evaluate Graduate Applications, GRADE, Allegedly Exacerbated Existing Inequality for Marginalized Applicants, Prompting Tool Suspension. *Artificial Intelligence Incident Database* (Dec. 2012). <https://incidentdatabase.ai/cite/135>
- [122] Patrick Hall. 2020. Incident 86: Coding Errors in Leaving Certificate Grading Algorithm Caused Inaccurate Scores in Ireland. *Artificial Intelligence Incident Database* (Oct. 2020). <https://incidentdatabase.ai/cite/86>
- [123] Justine Halley. 2021. The Siren Song of Exodus Cry:. <https://justinehalley.medium.com/the-siren-song-of-exodus-cry-d507a594c05d>
- [124] Karen Hao. 2019. AI Is Sending People to Jail—and Getting It Wrong. *MIT Technology Review* (Jan. 2019). <https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/>
- [125] Karen Hao. 2020. The Messy, Secretive Reality behind OpenAI's Bid to Save the World. *MIT Technology Review* (Feb. 2020). <https://www.technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality/>
- [126] Karen Hao and Nadine Freischlad. 2022. The Gig Workers Fighting Back against the Algorithms. *MIT Technology Review* (April 2022). <https://www.technologyreview.com/2022/04/21/1050381/the-gig-workers-fighting-back-against-the-algorithms/>
- [127] Karen Hao and Andrea Paola Hernández. 2022. How the AI Industry Profits from Catastrophe. *MIT Technology Review* (April 2022). <https://www.technologyreview.com/2022/04/20/1050392/ai-industry-appen-scale-data-labels/>
- [128] Karen Hao and Heidi Swart. 2022. South Africa's Private Surveillance Machine Is Fueling a Digital Apartheid. *MIT Technology Review* (April 2022). <https://www.technologyreview.com/2022/04/19/1049996/south-africa-ai-surveillance-digital-apartheid/>
- [129] Anna-Sophie Harling, Declan Henesy, and Eleanor Simmance. 2023. Transparency Reporting: The UK Regulatory Perspective. *Journal of Online Trust and Safety* 1, 5 (Jan. 2023). <https://doi.org/10.54501/jots.v1i5.108>
- [130] Mary Harris. 2021. Why Sex Workers Find the OnlyFans Fiasco So Familiar. *Slate* (Aug. 2021). <https://slate.com/technology/2021/08/onlyfans-porn-ban-reversal-sex-workers.html>
- [131] Rob Hawkins. 2023. About Reveddit: FAQ. <https://www.reveddit.com/about/faq/>
- [132] Will Douglas Heaven. 2023. The inside Story of How ChatGPT Was Built from the People Who Made It. *MIT Technology Review* (March 2023). <https://www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai/>
- [133] Simon Hegelich. 2020. Facebook Needs to Share More with Researchers. *Nature* 579, 7800 (March 2020), 473–473. <https://doi.org/10.1038/d41586-020-00828-5>
- [134] Simon Hegelich, Fabienne Marco, Joana Bayraktar, and Morteza Shahrezayee. 2020. The Social Science One Facebook Cooperation: A Systemic Failure. <https://politicaldatascience.blogspot.com/2020/03/the-social-science-one-facebook.html>
- [135] Alex Hern. 2021. Social Network Giants Pledge to Tackle Abuse of Women Online. *The Guardian* (July 2021). <https://www.theguardian.com/society/2021/jul/01/social-networks-facebook-google-twitter-tiktok-pledge-to-tackle-abuse-of-women-online>
- [136] Kashmir Hill. 2022. A Dad Took Photos of His Naked Toddler for the Doctor. Google Flagged Him as a Criminal. *The New York Times* (Aug. 2022). <https://www.nytimes.com/2022/08/21/technology/google-surveillance-toddler-photo.html>
- [137] Tarpley Hitt. 2020. Inside the Shady Sex-Work Abolitionist Group That Gutted Pornhub. *The Daily Beast* (Dec. 2020). <https://www.thedailybeast.com/inside-the-shady-sex-work-abolitionist-group-that-gutted-pornhub>
- [138] Tarpley Hitt. 2021. Facebook a Hotbed of 'Child Sexual Abuse Material'. *The Daily Beast* (Feb. 2021). <https://www.thedailybeast.com/facebook-a-hotbed-of-child-sexual-abuse-material-with-203-million-reports-far-more-than-pornhub>
- [139] Reid Hoffman and Chris Yeh. 2018. *Blitzscaling*. Currency, New York, NY, USA.
- [140] Amanda Hoover. 2023. The Mastodon Bump Is Now a Slump. *Wired* (Feb. 2023). <https://www.wired.com/story/the-mastodon-bump-is-now-a-slump/>
- [141] HRI. 2020. *The Global State of Harm Reduction 2020* (7 ed.). Harm Reduction International, London, United Kingdom. [https://hri.global/wp-content/uploads/2022/10/Global\\_State\\_HRI\\_2020\\_BOOK\\_FA\\_Web-1.pdf](https://hri.global/wp-content/uploads/2022/10/Global_State_HRI_2020_BOOK_FA_Web-1.pdf)
- [142] HRI. 2022. What is Harm Reduction? <https://hri.global/what-is-harm-reduction/>
- [143] Human Rights Council. 2018. *Report of the Independent International Fact-Finding Mission on Myanmar*. Mission Report A/HRC/39/64. Human Rights Council. [https://www.ohchr.org/sites/default/files/Documents/HRBodies/HRCouncil/FFM-Myanmar/A\\_HRC\\_39\\_64.pdf](https://www.ohchr.org/sites/default/files/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf)
- [144] Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. 2022. Robots Enact Malignant Stereotypes. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 743–756. <https://doi.org/10.1145/3531146.3533138>
- [145] Andrew Hutchinson. 2016. A Complete List of Facebook's Misreported Metrics and What They Mean. *Social Media Today* (Dec. 2016). <https://www.socialmediatoday.com/social-networks/complete-list-facebooks-misreported-metrics-and-what-they-mean>
- [146] Andrew Hutchinson. 2016. Facebook Clarifies Errors with Metrics - What It Means for Marketers. *Social Media Today* (Nov. 2016). <https://www.socialmediatoday.com/social-business/facebook-clarifies-errors-metrics-what-it-means-marketers?platform=hootsuite>

- [147] Andrew Hutchinson. 2016. On Facebook's Inflated Video Metrics and What It Means for Marketers. *Social Media Today* (Sept. 2016). <https://www.socialmediatoday.com/social-business/facebooks-inflated-video-metrics-and-what-it-means-marketers>
- [148] Andrew Hutchinson. 2017. Facebook Found Two New Errors in Their Ad Metrics, Issued Refunds. *Social Media Today* (Nov. 2017). <https://www.socialmediatoday.com/news/facebooks-found-two-new-errors-in-their-ad-metrics-issued-refunds/510291/>
- [149] Tomiwa Ilori. 2020. Content Moderation Is Particularly Hard in African Countries. *Slate* (Aug. 2020). <https://slate.com/technology/2020/08/social-media-content-moderation-african-nations.html>
- [150] Mathew Ingram. 2022. Facebook's New Data-Sharing Plans Raise Old Concerns. *Columbia Journalism Review* (May 2022). [https://www.cjr.org/the\\_media\\_today/facebooks-new-data-sharing-plans-raise-old-concerns.php](https://www.cjr.org/the_media_today/facebooks-new-data-sharing-plans-raise-old-concerns.php)
- [151] Instagram. 2021. Introducing Sensitive Content Control. <https://about.instagram.com/blog/announcements/introducing-sensitive-content-control>
- [152] Channarong Intahchomphoo and Odd Erik Gundersen. 2020. Artificial Intelligence and Race: A Systematic Review. *Legal Information Management* 20, 2 (June 2020), 74–84. <https://doi.org/10.1017/S1472669620000183>
- [153] Jeneen Interlandi. 2023. One Year Inside a Radical New Approach to America's Overdose Crisis. *The New York Times* (Feb. 2023). <https://www.nytimes.com/2023/02/22/opinion/drug-crisis-addiction-harm-reduction.html>
- [154] InvisibleDeck. 2022. "Jesus Christ Using His Cross as a Pogo Stick at Dawn in Judea; He Is in Midair; the Sun and the Sky Are Visible in the Background; Nokia 808 PureView; Motion Blur, 4k, Award-Winning Photograph". [www.reddit.com/r/dalle2/comments/w3lrg9/jesus\\_christ\\_using\\_his\\_cross\\_as\\_a\\_pogo\\_stick\\_at/](http://www.reddit.com/r/dalle2/comments/w3lrg9/jesus_christ_using_his_cross_as_a_pogo_stick_at/)
- [155] Rishi Iyengar. 2021. Facebook Has Language Blind Spots around the World That Allow Hate Speech to Flourish. *CNN* (Oct. 2021). <https://www.cnn.com/2021/10/26/tech/facebook-papers-language-hate-speech-international/index.html>
- [156] Jasper Jackson, Mark Townsend, and Lucy Kassa. 2022. Facebook 'Lets Vigilantes in Ethiopia Incite Ethnic Killing'. *The Observer* (Feb. 2022). <https://www.theguardian.com/technology/2022/feb/20/facebook-lets-vigilantes-in-ethiopia-incite-ethnic-killing>
- [157] Julian Jaurisch. 2022. Platform oversight: Here is what a strong Digital Services Coordinator should look like. *Verfassungsblog* (Oct. 2022). <https://doi.org/10.17176/20221031-220344-0>
- [158] Julian Jaurisch. 2022. *Platform Oversight: Here Is What a Strong Digital Services Coordinator Should Look Like*. Policy Brief. Stiftung Neue Verantwortung, Berlin, Germany. <https://www.stiftung-nv.de/en/publication/platform-oversight-what-strong-digital-services-coordinator-should-look>
- [159] Julian Jaurisch. 2023. Here Is Why Digital Services Coordinators Should Establish Strong Research and Data Units. <https://dsa-observatory.eu/2023/03/10/here-is-why-digital-services-coordinators-should-establish-strong-research-and-data-units/>
- [160] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. 2021. Understanding International Perceptions of the Severity of Harmful Content Online. *PLOS ONE* 16, 8 (Aug. 2021), e0256762. <https://doi.org/10.1371/journal.pone.0256762>
- [161] Jose L. Jimenez, Linsey C. Marr, Katherine Randall, Edward Thomas Ewing, Zeynep Tufekci, Trish Greenhalgh, Raymond Tellier, Julian W. Tang, Yuguo Li, Lidia Morawska, Jonathan Mesiano-Crookston, David Fisman, Orla Hegarty, Stephanie J. Dancer, Philomena M. Bluyssen, Giorgio Buonanno, Marcel G. L. C. Loomans, William P. Bahnfleth, Maosheng Yao, Chandra Sekhar, Pawel Wargocki, Arsen K. Melikov, and Kimberly A. Prather. 2022. What Were the Historical Reasons for the Resistance to Recognizing Airborne Transmission during the COVID-19 Pandemic? *Indoor Air* 32, 8 (Aug. 2022), e13070. <https://doi.org/10.1111/ina.13070>
- [162] Jolleb. 2022. "The Last Supper but in the Future". [www.reddit.com/r/dalle2/comments/wf9hcr/the\\_last\\_supper\\_but\\_in\\_the\\_future/](http://www.reddit.com/r/dalle2/comments/wf9hcr/the_last_supper_but_in_the_future/)
- [163] Franz Kafka. 1995. In the Penal Colony. In *The Complete Stories*. Knopf Doubleday Publishing Group, New York, NY, USA. <https://www.kafka-online.info/in-the-penal-colony.html>
- [164] Franz Kafka. 2005. *The Trial*. Project Gutenberg, Salt Lake City, UT, USA. <https://www.gutenberg.org/ebooks/7849/pg7849-images.html.utf8>
- [165] Jodi Kantor, Arya Sundaram, Aliza Aufrechtig, and Rumsey Taylor. 2022. The Rise of the Worker Productivity Score. *The New York Times* (Aug. 2022). <https://www.nytimes.com/interactive/2022/08/14/business/worker-productivity-tracking.html>
- [166] Jodi Kantor, Karen Weise, and Grace Ashford. 2021. The Amazon That Customers Don't See. *The New York Times* (June 2021). <https://www.nytimes.com/interactive/2021/06/15/us/amazon-workers.html>
- [167] Sayash Kapoor and Arvind Narayanan. 2023. A Misleading Open Letter about Sci-Fi AI Dangers Ignores the Real Risks. <https://aisnakeoil.substack.com/p/a-misleading-open-letter-about-sci>
- [168] Matthew Katsaros, Tom Tyler, Jisu Kim, and Tracey Meares. 2022. Procedural Justice and Self Governance on Twitter: Unpacking the Experience of Rule Breaking on Twitter. *Journal of Online Trust and Safety* 1, 3 (Aug. 2022). <https://doi.org/10.54501/jots.v1i3.38>
- [169] Sven Felix Kellerhoff. 2022. Spitzelstaat DDR: Die Stasi hatte viel mehr Informanten als bekannt. *Die Welt* (Jan. 2022). <https://www.welt.de/geschichte/article132502154/Spitzelstaat-DDR-Die-Stasi-hatte-viel-mehr-Informanten-als-bekannt.html>
- [170] Kate Klonick. 2018. The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review* 131, 6 (April 2018), 1598–1670. <https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/>
- [171] Lee Kovarsky. 2022. The Trump Executions. *Texas Law Review* 100, 4 (March 2022). <https://texaslawreview.org/the-trump-executions/>
- [172] Anastasia Kozyreva, Stefan M. Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. 2023. Resolving Content Moderation Dilemmas between Free Speech and Harmful Misinformation. *Proceedings of the National Academy of Sciences* 120, 7 (Feb. 2023), e2210666120. <https://doi.org/10.1073/pnas.2210666120>
- [173] Seth Kugel. 2022. Help! I Was Banned From Lyft and No One Will Tell Me Why. *The New York Times* (Nov. 2022). <https://www.nytimes.com/2022/11/17/travel/tripped-up-banned-lyft.html>

- [174] Tony Kushner. 1993. *Angels in America: A Gay Fantasia on National Themes, Part Two, Perestroika* (revised ed.). Theatre Communications Group, New York, NY, USA.
- [175] K. Hazel Kwon and Chun Shao. 2021. Dark Knowledge and Platform Governance: A Case of an Illicit E-Commerce Community in Reddit. *American Behavioral Scientist* 65, 6 (May 2021), 779–799. <https://doi.org/10.1177/0002764221989770>
- [176] Nick LaGrow and Miri Pruzan. 2019. Integrating Autoconversion: Facebook’s Path from Zawgyi to Unicode. <https://engineering.fb.com/2019/09/26/android/unicode-font-converter/>
- [177] Khoa Lam. 2020. Incident 374: UK Ofqual’s Algorithm Disproportionately Provided Lower Grades Than Teachers’ Assessments. *Artificial Intelligence Incident Database* (Aug. 2020). <https://incidentdatabase.ai/cite/374>
- [178] Tom Lauricella and Leslie Norton. 2021. How Facebook Silences Its Investors. *Morningstar* (Oct. 2021). <https://www.morningstar.com/articles/1061237/how-facebook-silences-its-investors>
- [179] Henry Charles Lea. 1906. *A History of the Inquisition of Spain*. Vol. 1. The Macmillan Company, New York, NY, USA. <http://archive.org/details/ahistoryinquisi03leagoog>
- [180] Henry Charles Lea. 1906. *A History of the Inquisition of Spain*. Vol. 2. The Macmillan Company, New York, NY, USA. <http://archive.org/details/historyofinquisispain02leah>
- [181] Henry Charles Lea. 1906. *A History of the Inquisition of Spain*. Vol. 4. The Macmillan Company, New York, NY, USA. <http://archive.org/details/historyofinquisi04leah>
- [182] Henry Charles Lea. 1906. *A History of the Inquisition of Spain*. Vol. 3. The Macmillan Company, New York, NY, USA. <http://archive.org/details/historyofinquisi03leahiala>
- [183] Alex Leavitt. 2015. “This Is a Throwaway Account”: Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW ’15)*. Association for Computing Machinery, New York, NY, USA, 317–327. <https://doi.org/10.1145/2675133.2675175>
- [184] Evelina Leivada, Elliot Murphy, and Gary Marcus. 2022. DALL-E 2 Fails to Reliably Capture Common Syntactic Processes. <https://doi.org/10.48550/arXiv.2210.12889> arXiv:arXiv:2210.12889
- [185] Natasha Lennard. 2020. Amazon Workers Are Organizing a Global Struggle. *The Intercept* (Dec. 2020). <https://theintercept.com/2020/12/03/amazon-workers-union-international-strike/>
- [186] Dyani Lewis. 2022. Why the WHO Took Two Years to Say COVID Is Airborne. *Nature* 604, 7904 (April 2022), 26–31. <https://doi.org/10.1038/d41586-022-00925-7>
- [187] Hsuan-Tien Lin, Maria Florina Balcan, Raia Hadsell, and Marc’Aurelio Ranzato. 2020. Announcing the NeurIPS 2020 Award Recipients. <https://neuripsconf.medium.com/announcing-the-neurips-2020-award-recipients-73e4d3101537>
- [188] Alfred Lua. 2022. 20 Top Social Media Sites to Consider for Your Brand in 2023. <https://buffer.com/library/social-media-sites/>
- [189] Ryan Mac and Craig Silverman. 2020. “Facebook Is Hurting People At Scale”: Mark Zuckerberg’s Employees Reckon With The Social Network They’ve Built. *BuzzFeed News* (July 2020). <https://www.buzzfeednews.com/article/ryanmac/facebook-employee-leaks-show-they-feel-betrayed>
- [190] Macrotrends. 2023. Alphabet: Number of Employees 2010-2022. <https://www.macrotrends.net/stocks/charts/GOOGL/alphabet/number-of-employees>
- [191] Macrotrends. 2023. Amazon: Number of Employees 2010-2022. <https://www.macrotrends.net/stocks/charts/AMZN/amazon/number-of-employees>
- [192] Macrotrends. 2023. Apple: Number of Employees 2010-2022. <https://www.macrotrends.net/stocks/charts/AAPL/apple/number-of-employees>
- [193] Macrotrends. 2023. Meta Platforms: Number of Employees 2010-2022. <https://www.macrotrends.net/stocks/charts/META/meta-platforms/number-of-employees>
- [194] Alexis C. Madrigal. 2019. Before It Conquered the World, Facebook Conquered Harvard. *The Atlantic* (Feb. 2019). <https://www.theatlantic.com/technology/archive/2019/02/and-then-there-was-thefacebookcom/582004/>
- [195] Odanga Madung. 2021. In Africa, Content Moderation Is a Dangerous Game. *Wired* (June 2021). <https://www.wired.com/story/opinion-in-africa-content-moderation-is-a-dangerous-game/>
- [196] Nesrine Malik. 2022. How Facebook Took over the Internet in Africa – and Changed Everything. *The Guardian* (Jan. 2022). <https://www.theguardian.com/technology/2022/jan/20/facebook-second-life-the-unstoppable-rise-of-the-tech-company-in-africa>
- [197] G. Alan Marlatt, Mary E Larimer, and Katie Witkiewitz (Eds.). 2011. *Harm Reduction: Pragmatic Strategies for Managing High-Risk Behaviors* (second ed.). Guilford Press, New York, NY, USA. <https://www.guilford.com/books/Harm-Reduction/Marlatt-Larimer-Witkiewitz/9781462502561>
- [198] Adam Martin. 2012. Goldman Sach’s Backpage Panic Is Real, Costly. <https://www.theatlantic.com/business/archive/2012/04/goldmans-backpage-panic-real-costly/329861/>
- [199] Thomas B. Marvell and Carlisle E. Moody. 2001. The Lethal Effects of Three-Strikes Laws. *The Journal of Legal Studies* 30, 1 (Jan. 2001), 89–106. <https://doi.org/10.1086/468112> jstor:10.1086/468112
- [200] Mike Masnick. 2017. Shockingly, NY Times Columnist Is Totally Clueless About The Internet. <https://www.techdirt.com/2017/09/19/shockingly-ny-times-columnist-is-totally-clueless-about-internet/>
- [201] Mike Masnick. 2022. Hey Elon: Let Me Help You Speed Run the Content Moderation Learning Curve. <https://www.techdirt.com/2022/11/02/hey-elon-let-me-help-you-speed-run-the-content-moderation-learning-curve/>
- [202] Adrienne Massanari. 2017. #Gamergate and The Fapping: How Reddit’s Algorithm, Governance, and Culture Support Toxic Technocultures. *New Media & Society* 19, 3 (March 2017), 329–346. <https://doi.org/10.1177/1461444815608807>

- [203] Nicholas McCormack. 2012. Nicholas Kristof Blasts Backpage For Aiding Child Prostitution. *HuffPost* (Jan. 2012). [https://www.huffpost.com/entry/nicholas-kristof-backpage\\_n\\_1234644](https://www.huffpost.com/entry/nicholas-kristof-backpage_n_1234644)
- [204] Erin McElroy, Meredith Whittaker, and Nicole E. Weber. 2021. Prison Tech Comes Home. <https://www.publicbooks.org/prison-tech-comes-home/>
- [205] Timothy McLaughlin. 2018. How Facebook’s Rise Fueled Chaos and Confusion in Myanmar. *Wired* (July 2018). <https://www.wired.com/story/how-facebooks-rise-fueled-chaos-and-confusion-in-myanmar/>
- [206] James McRitchie. 2020. Facebook 2020 Shareholders Request Equal Votes. <https://www.corpgov.net/2020/05/facebook-2020-shareholders-request-equal-votes/>
- [207] Jeremy B. Merrill and Ariana Tobin. 2019. Facebook Moves to Block Ad Transparency Tools — Including Ours. *ProPublica* (Jan. 2019). <https://www.propublica.org/article/facebook-blocks-ad-transparency-tools>
- [208] Meta Platforms. 2023. Meta Reports Fourth Quarter and Full Year 2022 Results. <https://investor.fb.com/investor-news/press-release-details/2023/Meta-Reports-Fourth-Quarter-and-Full-Year-2022-Results/default.aspx>
- [209] Cade Metz and Gregory Schmidt. 2023. Elon Musk and Others Call for Pause on A.I., Citing ‘Profound Risks to Society’. *The New York Times* (March 2023). <https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html>
- [210] Victoria Milko and Barbara Ortutay. 2022. ‘Kill More’: Facebook Fails to Detect Hate against Rohingya. <https://apnews.com/article/technology-business-bangladesh-myanmar-united-nations-f7d89e38c54f7bae464762fa23bd96b2>
- [211] Octave Mirbeau. 2008. *The Torture Garden*. Bookkake, London, United Kingdom. <https://www.bauerverlag.eu/downloads/torture-garden.pdf>
- [212] Pamela Mishkin and Lama Ahmad. 2022. DALL-E 2 Preview: Risks and Limitations. <https://github.com/openai/dalle-2-preview/blob/eeec5a1843b1d17cb9ed113117a2fcaa9206a564/system-card.md>
- [213] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT\* ’19)*. Association for Computing Machinery, New York, NY, USA, 220–229. <https://doi.org/10.1145/3287560.3287596>
- [214] Loveday Morris. 2021. In Poland’s Politics, a ‘Social Civil War’ Brewed as Facebook Rewarded Online Anger. *Washington Post* (Oct. 2021). <https://www.washingtonpost.com/world/2021/10/27/poland-facebook-algorithm/>
- [215] Paul Mozur. 2018. A Genocide Incited on Facebook, With Posts From Myanmar’s Military. *The New York Times* (Oct. 2018). <https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html>
- [216] Paul Mozur, Muyi Xiao, and John Liu. 2022. ‘An Invisible Cage’: How China Is Policing the Future. *The New York Times* (June 2022). <https://www.nytimes.com/2022/06/25/technology/china-surveillance-police.html>
- [217] Luke Munn. 2022. The Uselessness of AI Ethics. *AI and Ethics* (Aug. 2022). <https://doi.org/10.1007/s43681-022-00209-w>
- [218] Natalie. 2022. I Received a Warning While Using DALL-E 2. Will I Be Banned? <https://web.archive.org/web/20220818131242/https://help.openai.com/en/articles/6338765-i-received-a-warning-while-using-dall-e-2-will-i-be-banned>
- [219] NCMEC. 2020. *2019 CyberTipline Reports by Electronic Service Providers (ESP)*. Transparency Report. National Center for Missing and Exploited Children. <https://www.missingkids.org/content/dam/missingkids/pdfs/2019-reports-by-esp.pdf>
- [220] NCMEC. 2021. *2020 CyberTipline Reports by Electronic Service Providers (ESP)*. Transparency Report. National Center for Missing and Exploited Children. <https://www.missingkids.org/content/dam/missingkids/pdfs/2020-reports-by-esp.pdf>
- [221] NCMEC. 2022. *2021 CyberTipline Reports by Country*. Transparency Report. National Center for Missing and Exploited Children. <https://www.missingkids.org/content/dam/missingkids/pdfs/2021-reports-by-country.pdf>
- [222] NCMEC. 2022. *2021 Reports by Electronic Service Providers (ESP)*. Transparency Report. National Center for Missing and Exploited Children. <https://www.missingkids.org/content/dam/missingkids/pdfs/2021-reports-by-esp.pdf>
- [223] Ashley Nellis. 2023. *Mass Incarceration Trends*. Report. The Sentencing Project, Washington, DC. <https://www.sentencingproject.org/reports/mass-incarceration-trends/>
- [224] Deborah Netburn. 2021. A Timeline of the CDC’s Advice on Face Masks. *Los Angeles Times* (July 2021). <https://www.latimes.com/science/story/2021-07-27/timeline-cdc-mask-guidance-during-covid-19-pandemic>
- [225] Marissa Newman. 2023. Meta Was Scraping Sites for Years While Fighting the Practice. *Bloomberg* (Feb. 2023). <https://www.bloomberg.com/news/articles/2023-02-02/meta-was-scraping-sites-for-years-while-fighting-the-practice>
- [226] Casey Newton. 2019. Bodies in Seats. *The Verge* (June 2019). <https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa>
- [227] Casey Newton. 2019. The Secret Lives of Facebook Moderators in America. *The Verge* (Feb. 2019). <https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona>
- [228] Casey Newton. 2021. The Tier List: How Facebook Decides Which Countries Need Protection. *The Verge* (Oct. 2021). <https://www.theverge.com/22743753/facebook-tier-list-countries-leaked-documents-content-moderation>
- [229] Casey Newton. 2023. Microsoft Lays off Team That Taught Employees How to Make AI Tools Responsibly. *The Verge* (March 2023). <https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs>
- [230] Thoai D. Ngo. 2022. Bad COVID Public Health Messaging Is Blocking Our Path to a “New Normal”. *Scientific American* (June 2022). <https://www.scientificamerican.com/article/bad-covid-public-health-messaging-is-blocking-our-path-to-a-new-normal/>
- [231] Naomi Nix. 2021. Facebook Board Rejects Proposals to Curb Zuckerberg’s Power. *Bloomberg* (May 2021). <https://www.bloomberg.com/news/articles/2021-05-26/facebook-board-rejects-proposals-to-reduce-zuckerberg-s-power>



- [232] Amy O'Hara and Jodi Nelson. 2019. *Evaluation of the Social Science One – Social Science Research Council – Facebook Partnership*. Evaluation Report. William and Flora Hewlett Foundation. <https://www.hewlett.org/wp-content/uploads/2020/02/Facebook-Partnership-Final-Evaluation-Report.pdf>
- [233] Molly Olmstead. 2022. Why Did 11 Billion Alaskan Snow Crabs Suddenly Disappear? *Slate* (Oct. 2022). <https://slate.com/technology/2022/10/alaskan-snow-crabs-dead.html>
- [234] Open Society Foundations. 2021. Why Would You Give Clean Needles to Someone Who Uses Drugs? <https://www.opensocietyfoundations.org/explainers/what-harm-reduction>
- [235] OpenAI. 2018. OpenAI Charter. <https://openai.com/charter/>
- [236] OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. <https://openai.com/blog/chatgpt/>
- [237] OpenAI. 2022. DALL·E API Now Available in Public Beta. <https://openai.com/blog/dall-e-api-now-available-in-public-beta/>
- [238] OpenAI. 2022. DALL·E Now Available Without Waitlist. <https://openai.com/blog/dall-e-now-available-without-waitlist/>
- [239] OpenAI. 2022. New-and-Improved Content Moderation Tooling. <https://openai.com/blog/new-and-improved-content-moderation-tooling/>
- [240] OpenAI. 2022. Reducing Bias and Improving Safety in DALL·E 2. <https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/>
- [241] OpenAI. 2022. Updates From OpenAI.
- [242] Barbara Ortutay. 2022. Amnesty Report Finds Facebook Amplified Hate Ahead of Rohingya Massacre in Myanmar. <https://www.pbs.org/newshour/world/amnesty-report-finds-facebook-amplified-hate-ahead-of-rohingya-massacre-in-myanmar>
- [243] George Orwell. 1949. *Nineteen Eighty-Four*. Secker & Warburg, London, United Kingdom.
- [244] Nilay Patel. 2022. How to Buy a Social Network, with Tumblr CEO Matt Mullenweg. *The Verge* (Dec. 2022). <https://www.theverge.com/23506085/wordpress-twitter-tumblr-ceo-matt-mullenweg-elon-musk>
- [245] Nilay Patel. 2022. Welcome to Hell, Elon. *The Verge* (Oct. 2022). <https://www.theverge.com/2022/10/28/23428132/elon-musk-twitter-acquisition-problems-speech-moderation>
- [246] Nilay Patel. 2023. Can Mastodon Seize the Moment from Twitter? *The Verge* (March 2023). <https://www.theverge.com/23658648/mastodon-ceo-twitter-interview-elon-musk-twitter>
- [247] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 19th International Conference on Supporting Group Work (GROUP '16)*. Association for Computing Machinery, New York, NY, USA, 369–374. <https://doi.org/10.1145/2957276.2957297>
- [248] Jeff Patton and Peter Economy. 2014. *User Story Mapping*. O'Reilly Media, Sebastopol, CA, USA. <https://www.jpattonassociates.com/jeff-pattons-book-released-user-story-mapping/>
- [249] Umashanthi Pavalanathan and Munmun De Choudhury. 2015. Identity Management and Mental Health Discourse in Social Media. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion)*. Association for Computing Machinery, New York, NY, USA, 315–321. <https://doi.org/10.1145/2740908.2743049>
- [250] Billy Perrigo. 2022. Inside Facebook's African Sweatshop. *Time* (Feb. 2022). <https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/>
- [251] Billy Perrigo. 2023. Facebook's 'Ethical' Outsourcing Partner Quits Content Moderation Work. *Time* (Jan. 2023). <https://time.com/6246018/facebook-sama-quits-content-moderation/>
- [252] Billy Perrigo. 2023. OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic. *Time* (Jan. 2023). <https://time.com/6247678/openai-chatgpt-kenya-workers/>
- [253] Walter L. Perry, Brian McInnis, Carter C. Price, Susan Smith, and John S. Hollywood. 2013. *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. RAND Corporation, Santa Monica, CA. <https://doi.org/10.7249/RR233>
- [254] Jay Peters. 2022. More than Two Million Users Have Flocked to Mastodon since Elon Musk Took over Twitter. *The Verge* (Dec. 2022). <https://www.theverge.com/2022/12/20/23518325/mastodon-monthly-active-users-twitter-elon-musk>
- [255] Pinterest. 2022. *Q1+Q2 2022 Transparency Report*. Transparency Report Q1+Q2 2022. Pinterest. <https://policy.pinterest.com/en/transparency-report>
- [256] Chavez Procope, Adeel Cheema, David Adkins, Bilal Alsallakh, Nekesha Green, Emily McReynolds, Grace Pehl, Erin Wang, and Polina Zvyagina. 2022. *System-Level Transparency of Machine Learning*. Technical Report. Meta. <https://ai.facebook.com/research/publications/system-level-transparency-of-machine-learning>
- [257] Producer X. 2020. Porn vs Pulitzer: How Nicholas Kristof and the New York Times Became Mouthpieces for Religious.... <https://producerx.medium.com/porn-vs-pulitzer-how-nicholas-kristof-and-the-new-york-times-became-mouthpieces-for-religious-be8db923ad07>
- [258] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society* 7, 2 (April 2021), 20563051211019004. <https://doi.org/10.1177/20563051211019004>
- [259] Carina Prunkl, Carolyn Ashurst, Markus Anderljung, Helena Webb, Jan Leike, and Allan Dafoe. 2021. Institutionalising Ethics in AI through Broader Impact Requirements. *Nature Machine Intelligence* 3, 2 (Feb. 2021), 104–110. <https://doi.org/10.1038/s42256-021-00298-y> arXiv:2106.11039 [cs]
- [260] Aneesh Raghunandan and Shivaram Rajgopal. 2022. Do Socially Responsible Firms Walk the Talk? <https://doi.org/10.2139/ssrn.3609056>
- [261] Abdul Rahman Al Jaloud, Hadi Al Khatib, Jeff Deutch, Dia Kayyali, and Jillian C. York. 2019. *Caught in the Net: The Impact of "Extremist" Speech Regulations on Human Rights Content*. White Paper. Electronic Frontier Foundation, Syrian Archive, and Witness. <https://www.eff.org/wp/caught-net-impact-extremist-speech-regulations-human-rights-content>
- [262] Inioluwa Deborah Raji, Timnit Gebbru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for

- Computing Machinery, New York, NY, USA, 145–151. <https://doi.org/10.1145/3375627.3375820>
- [263] Aravindh Raman, Sagar Joglekar, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2019. Challenges in the Decentralised Web: The Mastodon Case. In *Proceedings of the 2019 ACM Internet Measurement Conference*. Association for Computing Machinery, Amsterdam, Netherlands, 217–229. <https://doi.org/10.1145/3355369.3355572> arXiv:1909.05801 [cs]
- [264] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. <https://doi.org/10.48550/arXiv.2204.06125> arXiv:arXiv:2204.06125
- [265] James Redden, Molly O'Donovan Dix, and Criminal Justice Testing and Evaluation Consortium. 2020. Artificial Intelligence in the Criminal Justice System. <https://cjtec.org/artificial-intelligence-in-the-criminal-justice-system/>
- [266] Peter Redfield. 2005. Foucault in the Tropics: Displacing the Panopticon. In *Anthropologies of Modernity: Foucault, Governmentality, and Life Politics*, Jonathan Xavier Inda (Ed.). Blackwell Publishing, Hoboken, NJ, USA, 50–79. <https://redfield.web.unc.edu/wp-content/uploads/sites/9305/2015/08/foucault-tropics.pdf>
- [267] Eric Reinhart. 2022. How Joe Biden Launched a New Prison Boom. *Slate* (Jan. 2022). <https://slate.com/news-and-politics/2022/01/how-president-biden-launched-a-new-prison-boom.html>
- [268] René Belbenoit. 1938. *Dry Guillotine: Fifteen Years among the Living Dead*. Blue Ribbon Books, New York, NY, USA. [http://archive.org/details/Dry\\_Guillotine](http://archive.org/details/Dry_Guillotine)
- [269] Diana Rieger, Anna Sophie Kümpel, Maximilian Wich, Toni Kiening, and Georg Groh. 2021. Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit. *Social Media + Society* 7, 4 (Oct. 2021), 20563051211052906. <https://doi.org/10.1177/20563051211052906>
- [270] Ritchie Robertson. 2017. Kafka's Reading. In *Franz Kafka in Context*, Carolin Duttlinger (Ed.). Cambridge University Press, Cambridge, 82–90. <https://doi.org/10.1017/9781316084243.011>
- [271] Nick Robins-Early. 2021. How Facebook Is Stoking a Civil War in Ethiopia. <https://www.vice.com/en/article/qjbpd7/how-facebook-is-stoking-a-civil-war-in-ethiopia>
- [272] Kevin Roose. 2021. Inside Facebook's Data Wars. *The New York Times* (July 2021). <https://www.nytimes.com/2021/07/14/technology/facebook-data.html>
- [273] Alex Rosenblat. 2018. *Uberland: How Algorithms Are Rewriting the Rules of Work* (first ed.). University of California Press, Berkeley, CA, USA. jstor:10.1525/j.ctv5cgbm3 <https://www.jstor.org/stable/10.1525/j.ctv5cgbm3>
- [274] Yoel Roth and Vijaya Gadde. 2021. Expanding Access beyond Information Operations. [https://blog.twitter.com/en\\_us/topics/company/2021-expanding-access-beyond-information-operations-](https://blog.twitter.com/en_us/topics/company/2021-expanding-access-beyond-information-operations-)
- [275] Alan Rozenshtein. 2022. Mastodon's Content-Moderation Growing Pains. *Reason* (Nov. 2022). <https://reason.com/volokh/2022/11/21/mastodons-content-moderation-growing-pains/>
- [276] Michael Sainato. 2021. 'I'm Still in Pain': Amazon Employees Say Climate of Fear Has Led to High Rates of Injuries. *The Guardian* (Dec. 2021). <https://www.theguardian.com/technology/2021/dec/30/amazon-employees-climate-fear-high-rates-injuries>
- [277] Michael Sainato. 2022. Amazon Could Run out of Workers in US in Two Years, Internal Memo Suggests. *The Guardian* (June 2022). <https://www.theguardian.com/technology/2022/jun/22/amazon-workers-shortage-leaked-memo-warehouse>
- [278] Vishwanath Sarang. 2022. Community Standards Enforcement Report Assessment Results. <https://about.fb.com/news/2022/05/community-standards-enforcement-report-assessment-results/>
- [279] Adam Satariano. 2021. After Barring Trump, Facebook and Twitter Face Scrutiny About Inaction Abroad. *The New York Times* (Jan. 2021). <https://www.nytimes.com/2021/01/14/technology/trump-facebook-twitter.html>
- [280] Wendy Sawyer and Peter Wagner. 2022. Mass Incarceration: The Whole Pie 2022. <https://www.prisonpolicy.org/reports/pie2022.html>
- [281] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A Framework of Severity for Harmful Content Online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 368:1–368:33. <https://doi.org/10.1145/3479512>
- [282] Zoë Schiffer, Casey Newton, and Alex Heath. 2023. Inside Elon's 'Extremely Hardcore' Twitter. *New York Magazine* (Jan. 2023). <https://nymag.com/intelligencer/article/elon-musk-twitter-takeover.html>
- [283] Ari Schlesinger, Kenton P. O'Hara, and Alex S. Taylor. 2018. Let's Talk About Race: Identity, Chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–14. <https://doi.org/10.1145/3173574.3173889>
- [284] Angela M. Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. 2022. Why Do Volunteer Content Moderators Quit? Burnout, Conflict, and Harmful Behaviors. *New Media & Society* (Dec. 2022), 14614448221138529. <https://doi.org/10.1177/14614448221138529>
- [285] Nicolas Schradang, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. 2015. An Analysis of Domestic Abuse Discourse on Reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 2577–2583. <https://doi.org/10.18653/v1/D15-1309>
- [286] Klaus Schroeder. 2013. *Der SED-Staat: Geschichte Und Strukturen Der DDR 1949-1990* (third ed.). Böhlau Verlag, Wien, Austria and Köln, Germany. <https://www.vandenhoeck-ruprecht-verlage.com/themen-entdecken/geschichte/geschichte-der-neuzeit/42210/der-sed-staat>
- [287] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. <https://doi.org/10.48550/arXiv.2111.02114> arXiv:arXiv:2111.02114



- [288] Dylan Scott. 2022. The Most Consistently Botched Part of the US Pandemic Response. *Vox* (Jan. 2022). <https://www.vox.com/coronavirus-covid19/22870268/cdc-covid-19-guidelines-isolation-boosters-masks>
- [289] Mark Scott. 2021. Facebook Did Little to Moderate Posts in the World's Most Violent Countries. *POLITICO* (Oct. 2021). <https://www.politico.com/news/2021/10/25/facebook-moderate-posts-violent-countries-517050>
- [290] Mark Scott and Laura Kayali. 2020. What Happened When Humans Stopped Managing Social Media Content. <https://www.politico.eu/article/facebook-content-moderation-automation/>
- [291] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator Engagement and Community Development in the Age of Algorithms. *New Media & Society* 21, 7 (July 2019), 1417–1443. <https://doi.org/10.1177/1461444818821316>
- [292] Serious\_Historian578. 2022. Dalle "AI Alignment" by Randomly Adding Words to the Ends of Prompts. [www.reddit.com/r/slatestarcode/comments/w5bjrz/dalle\\_ai\\_alignment\\_by\\_randomly\\_adding\\_words\\_to/](https://www.reddit.com/r/slatestarcode/comments/w5bjrz/dalle_ai_alignment_by_randomly_adding_words_to/)
- [293] Riddhi Setty. 2023. Getty Images Sues Stability AI Over Art Generator IP Violations. *Bloomberg Law* (Feb. 2023). <https://news.bloomberglaw.com/ip-law/getty-images-sues-stability-ai-over-art-generator-ip-violations>
- [294] Emily Shapiro and Meredith Deliso. 2022. Parkland Sentencing: Nikolas Cruz Sentenced to Life in Prison. *ABC News* (Nov. 2022). <https://abcnews.go.com/US/live-updates/parkland-nikolas-cruz/?id=91883130>
- [295] Alex Sherman. 2021. Reddit's CEO Has a Colorful Nickname for the Redditors Who Ruin It for Everyone. *CNBC* (Jan. 2021). <https://www.cnn.com/2021/01/29/reddit-ceo-has-colorful-nickname-for-troublemakers.html>
- [296] Craig Silverman and Bianca Fortis. 2023. A Scammer Who Tricks Instagram Into Banning Influencers Has Never Been Identified. We May Have Found Him. *ProPublica* (March 2023). <https://www.propublica.com/article/instagram-fraudster-ban-influencer-accounts>
- [297] Craig Silverman, Ryan Mac, and Pranav Dixit. 2020. "I Have Blood On My Hands": A Whistleblower Says Facebook Ignored Global Political Manipulation. *BuzzFeed News* (Sept. 2020). <https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo>
- [298] Jonathan Simon. 2007. Rise of the Carceral State. *Social Research* 74, 2 (2007), 471–508. jstor:40971941 <https://www.jstor.org/stable/40971941>
- [299] Tom Simonite. 2021. Facebook Is Everywhere; Its Moderation Is Nowhere Close. *Wired* (Oct. 2021). <https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/>
- [300] Sudhvir Singh, Christine McNab, Rose McKeon Olson, Nellie Bristol, Cody Nolan, Elin Bergström, Michael Bartos, Shunsuke Mabuchi, Raj Panjabi, Abraar Karan, Salma M. Abdalla, Mathias Bonk, Margaret Jamieson, George K. Werner, Anders Nordström, Helena Legido-Quigley, and Alexandra Phelan. 2021. How an Outbreak Became a Pandemic: A Chronological Analysis of Crucial Junctures and International Obligations in the Early Months of the COVID-19 Pandemic. *The Lancet* 398, 10316 (Nov. 2021), 2109–2124. [https://doi.org/10.1016/S0140-6736\(21\)01897-3](https://doi.org/10.1016/S0140-6736(21)01897-3)
- [301] Ellen Johnson Sirleaf and Helen Clark. 2021. Report of the Independent Panel for Pandemic Preparedness and Response: Making COVID-19 the Last Pandemic. *The Lancet* 398, 10295 (May 2021), 101–103. [https://doi.org/10.1016/S0140-6736\(21\)01095-3](https://doi.org/10.1016/S0140-6736(21)01095-3)
- [302] Ellen Johnson Sirleaf, Helen Clark, Mauricio Cárdenas, Aya Chebbi, Mark Dybul, Michel Kazatchkine, Joanne Liu, Precious Matsoso, David Miliband, Thoraya Obaid, Preeti Sudan, Zhong Nanshan, and Ernesto Zedillo. 2021. *COVID-19: Make It the Last Pandemic*. Technical Report. The Independent Panel for Pandemic Preparedness and Response. <https://theindependentpanel.org/mainreport/>
- [303] Jack Smith IV. 2016. China Is Creating a Thoughtcrime-Predicting Surveillance System for Its Citizens. <https://www.mic.com/articles/137464/china-is-creating-a-thoughtcrime-predicting-surveillance-system-for-its-citizens>
- [304] Matthew Sparkes. 2022. AI Art Tool DALL-E 2 Adds 'black' or 'Female' to Some Image Prompts. *New Scientist* (July 2022). <https://www.newscientist.com/article/2329690-ai-art-tool-dall-e-2-adds-black-or-female-to-some-image-prompts/>
- [305] Spicy-Elephant. 2022. My Friends and I Got Banned Today. I Said "Shrek with Red Eyes, Head Shot" as in a Camera Shot Focused on His Head. It Thought I Meant Shooting His Head. So Long Dall e. [www.reddit.com/r/dalle2/comments/wisk3y/my\\_friends\\_and\\_i\\_got\\_banned\\_today\\_i\\_said\\_shrek/](https://www.reddit.com/r/dalle2/comments/wisk3y/my_friends_and_i_got_banned_today_i_said_shrek/)
- [306] Pieter Spierenburg. 2009. Stephen A. Toth, Beyond Papillon. The French Overseas Penal Colonies, 1854-1952. *Crime, Histoire & Sociétés / Crime, History & Societies* 13, 1 (March 2009), 153–155. <https://journals.openedition.org/chs/716>
- [307] Daniel Sprick. 2019. Predictive Policing in China: An Authoritarian Dream of Public Security. *NAVEIN REET: Nordic Journal of Law & Social Research* 1, 9 (2019), 299–324. <https://tidsskrift.dk/njlsr/article/view/122164/169413>
- [308] Ramya Srinivasan and Ajay Chander. 2021. Biases in AI Systems. *Commun. ACM* 64, 8 (July 2021), 44–49. <https://doi.org/10.1145/3464903>
- [309] Stability AI. 2022. Stable Diffusion Public Release. <https://stability.ai/blog/stable-diffusion-public-release>
- [310] Nick Stockton. 2020. Incident 78: Meet the Secret Algorithm That's Keeping Students Out of College. *Artificial Intelligence Incident Database* (July 2020). <https://incidentdatabase.ai/cite/78>
- [311] Sheryl Gay Stolberg and Astead W. Herndon. 2019. 'Lock the S.O.B.s Up': Joe Biden and the Era of Mass Incarceration. *The New York Times* (June 2019). <https://www.nytimes.com/2019/06/25/us/joe-biden-crime-laws.html>
- [312] Jessica Stoya. 2021. What We Can Really Learn From the OnlyFans Debacle. *Slate* (Aug. 2021). <https://slate.com/human-interest/2021/08/onlyfans-sex-banned-allowed-decision-history.html>
- [313] Peter Strang, Jonas Bergström, Lisa Martinsson, and Staffan Lundström. 2020. Dying From COVID-19: Loneliness, End-of-Life Discussions, and Support for Patients and Their Families in Nursing Homes and Hospitals. A National Register Study. *Journal of Pain and Symptom Management* 60, 4 (Oct. 2020), e2–e13. <https://doi.org/10.1016/j.jpainsymman.2020.07.020>

- [314] Mathilda Strom. 2016. Why Does Facebook = the Internet in Myanmar? <https://www.linkedin.com/pulse/why-does-facebook-internet-myanmar-mathilda-strom>
- [315] Students at the University of Bristol. 2021. The Dreyfus Affair and the Image of the Intellectual. <https://intellectualsandthemedias.org/2021/12/11/the-dreyfus-affair-and-the-image-of-the-intellectual/>
- [316] Asawin Suebsaeng and Patrick Reis. 2023. Trump's Killing Spree: The Inside Story of His Race to Execute Every Prisoner He Could. *Rolling Stone* (Jan. 2023). <https://www.rollingstone.com/politics/politics-features/trump-capital-punishment-brandon-bernard-lisa-montgomery-1234664126/>
- [317] Levi Sumagaysay. 2019. Facebook Shareholders Fail to Rein in Mark Zuckerberg. *The Mercury News* (May 2019). <https://www.mercurynews.com/2019/05/30/ahead-of-todays-shareholders-meeting-calls-to-fire-zuck-and-break-up-facebook/>
- [318] G Supran and S Rahmstorf. 2023. Assessing ExxonMobil's Global Warming Projections. *Science* 379, 6628 (Jan. 2023). <https://doi.org/10.1126/science.abk0063>
- [319] Hiroko Tabuchi. 2021. In Your Facebook Feed: Oil Industry Pushback Against Biden Climate Plans. *The New York Times* (Sept. 2021). <https://www.nytimes.com/2021/09/30/climate/api-exxon-biden-climate-bill.html>
- [320] Meredith Talusan. 2017. How the Shutdown of Backpage Disproportionately Affects Trans Sex Workers. <https://www.vice.com/en/article/wnnedw/how-the-shutdown-of-backpage-disproportionately-affects-trans-sex-workers>
- [321] Matthew Taylor and Jonathan Watts. 2019. Revealed: The 20 Firms behind a Third of All Carbon Emissions. *The Guardian* (Oct. 2019). <https://www.theguardian.com/environment/2019/oct/09/revealed-20-firms-third-carbon-emissions>
- [322] Pim ten Thije. 2022. European Commission Articulates Priorities for Implementing the DSA - DSA Observatory. <https://dsa-observatory.eu/2022/11/01/european-commission-priorities-implementing-delegated-guidelines-acts-dsa-digital-services-act/>
- [323] TheDrewDude. 2022. "Jesus Christ Taking a Selfie While on a Cross". [www.reddit.com/r/dalle2/comments/wkmgz5/jesus\\_christ\\_taking\\_a\\_selfie\\_while\\_on\\_a\\_cross/](http://www.reddit.com/r/dalle2/comments/wkmgz5/jesus_christ_taking_a_selfie_while_on_a_cross/)
- [324] Nitasha Tiku. 2018. After Troubles in Myanmar, Facebook Charges Ahead in Africa. *Wired* (Oct. 2018). <https://www.wired.com/story/after-troubles-myanmar-facebook-charges-ahead-africa/>
- [325] Craig Timberg. 2021. Facebook Made Big Mistake in Data It Provided to Researchers, Undermining Academic Work. *Washington Post* (Sept. 2021). <https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/>
- [326] Twitter. 2021. *Rules Enforcement*. Transparency Report. Twitter. <https://transparency.twitter.com/en/reports/rules-enforcement.html>
- [327] Tom R. Tyler. 2003. Procedural Justice, Legitimacy, and the Effective Rule of Law. *Crime and Justice* 30 (2003), 283–357. jstor:1147701 <https://www.jstor.org/stable/1147701>
- [328] Tom R Tyler. 2007. Procedural Justice and the Courts. *Court Review* 44, 1–2 (2007), 26–31. <https://amjudges.org/publications/courtrv/cr44-1/CR44-1-2Tyler.pdf>
- [329] Tom R. Tyler. Sun, 05/07/2006 - 12:00. *Why People Obey the Law*. Princeton University Press, Princeton, NJ, USA. <https://press.princeton.edu/books/paperback/9780691126739/why-people-obey-the-law>
- [330] UNICEF. 2023. *Ethiopia: January - December 2022*. Humanitarian Situation Report 12. UNICEF. <https://reliefweb.int/report/ethiopia/unicef-ethiopia-humanitarian-situation-report-no-12-january-december-2022>
- [331] Jerry Useem. 2020. Beware of Corporate Promises. *The Atlantic* (Aug. 2020). <https://www.theatlantic.com/ideas/archive/2020/08/companies-stand-solidarity-are-licensing-themselves-discriminate/614947/>
- [332] Mathias Vermeulen. 2022. Researcher Access to Platform Data: European Developments. *Journal of Online Trust and Safety* 1, 4 (Sept. 2022). <https://doi.org/10.54501/jots.v1i4.84>
- [333] James Vincent. 2023. OpenAI Co-Founder on Company's Past Approach to Openly Sharing Research: "We Were Wrong". *The Verge* (March 2023). <https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview>
- [334] Michael Vitiello. 2002. Three Strikes Laws: A Real or Imagined Deterrent to Crime? *Human Rights Magazine* 29, 3 (April 2002). [https://www.americanbar.org/groups/crsj/publications/human\\_rights\\_magazine\\_home/human\\_rights\\_vol29\\_2002/spring2002/hr\\_spring02\\_vitiello/](https://www.americanbar.org/groups/crsj/publications/human_rights_magazine_home/human_rights_vol29_2002/spring2002/hr_spring02_vitiello/)
- [335] Suzanne Vranica and Jack Marshall. 2016. Facebook Overestimated Key Video Metric for Two Years. *Wall Street Journal* (Sept. 2016). <https://www.wsj.com/articles/facebook-overestimated-key-video-metric-for-two-years-1474586951>
- [336] Author Olin Wade. 2022. How Do I Read My Burmese Font on Facebook? [www remodelormove.com/how-do-i-read-my-burmese-font-on-facebook/](http://www remodelormove.com/how-do-i-read-my-burmese-font-on-facebook/)
- [337] Kurt Wagner. 2015. Facebook Is Opening Its First Office in Africa. *Vox* (June 2015). <https://www.vox.com/2015/6/29/11563948/facebook-is-opening-its-first-office-in-africa>
- [338] Peter Wagner and Daniel Kopf. 2015. The Racial Geography of Mass Incarceration. <https://www.prisonpolicy.org/racialgeography/report.html>
- [339] Glenn K. Wakam, John R. Montgomery, Ben E. Biesterveld, and Craig S. Brown. 2020. Not Dying Alone — Modern Compassionate Care in the Covid-19 Pandemic. *New England Journal of Medicine* 382, 24 (June 2020), e88. <https://doi.org/10.1056/NEJMp2007781>
- [340] David Wallechinsky and Irving Wallace. 1978. Famous Lasts: The Last Prisoner on Devil's Island. <https://archive.org/details/peoplesalmanac200wall>
- [341] Wasted-Entity. 2022. "Jesus Christ Wielding a Samurai Sword and Riding on the Back of a Velociraptor, Painting". [www.reddit.com/r/dalle2/comments/v2bb19/jesus\\_christ\\_wielding\\_a\\_samurai\\_sword\\_and\\_riding/](http://www.reddit.com/r/dalle2/comments/v2bb19/jesus_christ_wielding_a_samurai_sword_and_riding/)
- [342] John Waters. 1988. Hairspray.
- [343] Richard Waters and Harriet Agnew. 2022. Meta Shareholders Vent Anger at Zuckerberg's Spending Binge. *Financial Times* (Oct. 2022). <https://www.ft.com/content/0f4c676c-56a6-4b5e-850f-ddb78f9feb40>

- [344] Elizabeth Weil. 2023. You Are Not a Parrot. *Intelligencer* (March 2023). <https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html>
- [345] Bari Weiss, Abigail Shrier, Michael Shellenberger, and Nellie Bowles. 2022. Twitter’s Secret Blacklists. *The Free Press* (Dec. 2022). <https://www.thefp.com/p/twitters-secret-blacklists>
- [346] Jess Whittlestone, Rune Nyrop, Anna Alexandrova, and Stephen Cave. 2019. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’19)*. Association for Computing Machinery, New York, NY, USA, 195–200. <https://doi.org/10.1145/3306618.3314289>
- [347] Wikipedia. 2023. Francis Bacon (Artist). *Wikipedia* (Jan. 2023). [https://en.wikipedia.org/wiki/Francis\\_Bacon\\_\(artist\)](https://en.wikipedia.org/wiki/Francis_Bacon_(artist))
- [348] Wikipedia editors. 2023. The Blitz. *Wikipedia* (March 2023). [https://en.wikipedia.org/w/index.php?title=The\\_Blitz&oldid=1147335624](https://en.wikipedia.org/w/index.php?title=The_Blitz&oldid=1147335624)
- [349] Belkis Wille. 2020. “Video Unavailable”. Report. Human Rights Watch. <https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes>
- [350] Julia Carrie Wong and Hannah Ellis-Petersen. 2021. Facebook Planned to Remove Fake Accounts in India – until It Realized a BJP Politician Was Involved. *The Guardian* (April 2021). <https://www.theguardian.com/technology/2021/apr/15/facebook-india-bjp-fake-accounts>
- [351] Julia Carrie Wong and Jeff Ernst. 2021. Facebook Knew of Honduran President’s Manipulation Campaign – and Let It Continue for 11 Months. *The Guardian* (April 2021). <https://www.theguardian.com/technology/2021/apr/13/facebook-honduras-juan-orlando-hernandez-fake-engagement>
- [352] Julia Carrie Wong and Luke Harding. 2021. ‘Facebook Isn’t Interested in Countries like Ours’: Azerbaijan Troll Network Returns Months after Ban. *The Guardian* (April 2021). <https://www.theguardian.com/technology/2021/apr/13/facebook-azerbaijan-ilham-aliyev>
- [353] World Prison Brief. 2023. Highest to Lowest - Prison Population Rate. [https://www.prisonstudies.org/highest-to-lowest/prison\\_population\\_rate?field\\_region\\_taxonomy\\_tid=All](https://www.prisonstudies.org/highest-to-lowest/prison_population_rate?field_region_taxonomy_tid=All)
- [354] Worldometer. 2023. United States Population. <https://www.worldometers.info/world-population/us-population/>
- [355] Roman Yampolskiy. 2015. Incident 57: Australian Automated Debt Assessment System Issued False Notices to Thousands. *Artificial Intelligence Incident Database* (July 2015). <https://incidentdatabase.ai/cite/57/>
- [356] Roman Yampolskiy. 2016. Incident 11: Northpointe Risk Models. *Artificial Intelligence Incident Database* (May 2016). <https://incidentdatabase.ai/cite/11>
- [357] Jisheng Yang. 2021. *The World Turned Upside Down: A History of the Chinese Cultural Revolution*. Farrar, Straus and Giroux, New York, NY, USA. <https://us.macmillan.com/books/9781250829702/theworldturnedupsideup>
- [358] Eliezer Yudkowsky. 2023. The Open Letter on AI Doesn’t Go Far Enough. *Time* (March 2023). <https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/>
- [359] Cat Zahrzewski, Gerrit De Vynck, Niha Masihi, and Shibani Mahtani. 2021. How Facebook Neglected the Rest of the World, Fueling Hate Speech and Violence in India. *Washington Post* (Oct. 2021). <https://www.washingtonpost.com/technology/2021/10/24/india-facebook-misinformation-hate-speech/>
- [360] Zecharias Zelalem and Peter Guest. 2021. Why Facebook Keeps Failing in Ethiopia. *Rest of World* (Nov. 2021). <https://restofworld.org/2021/why-facebook-keeps-failing-in-ethiopia/>
- [361] David Zweig. 2022. How Twitter Rigged the Covid Debate. *The Free Press* (Dec. 2022). <https://www.thefp.com/p/how-twitter-rigged-the-covid-debate>

## A POLICY AND TERMS OF USE FOR DALL•E 2

Section A.1 documents DALL•E 2’s content policy and Section A.2 its addendum to OpenAI’s terms of use, both as of July 20, 2022. They preserve the structure of the original, with links pointing to pages in the Internet Archive. Note that archived OpenAI webpages may contain CSS that prevents the printing of a full page and JavaScript that redirects to an error page after a few seconds.

The content policy still is located at <https://labs.openai.com/policies/content-policy>. It was updated on September 19, 2022 by rewording the rules on disclosing the role of AI and by removing the fourth bullet of the rules on respecting the rights of others. At that time, OpenAI also updated its notification for violative prompts to state “It looks like this request may not follow our content policy.” above the cartoon shown in Fig. 1. Given this abrupt switch from the inappropriately punitive to the inappropriately saccharine, one wonders whether OpenAI sprung Dolores Umbridge out of Azkaban prison and tasked her with content policy enforcement.

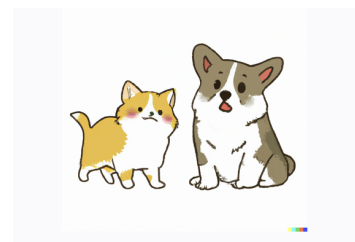


Fig. 1. Be nice to DALL•E’s pets!

The addendum to OpenAI's terms of use was located at <https://labs.openai.com/policies/terms>, but was rescinded on November 4, 2022.

## A.1 Content Policy

Thank you for trying our generative AI tools!

In your usage, you must adhere to our Content Policy:

**Do not attempt to create, upload, or share images that are not G-rated or that could cause harm.**

- **Hate:** hateful symbols, negative stereotypes, comparing certain groups to animals/objects, or otherwise expressing or promoting hate based on identity.
- **Harassment:** mocking, threatening, or bullying an individual.
- **Violence:** violent acts and the suffering or humiliation of others.
- **Self-harm:** suicide, cutting, eating disorders, and other attempts at harming oneself.
- **Sexual:** nudity, sexual acts, sexual services, or content otherwise meant to arouse sexual excitement.
- **Shocking:** bodily fluids, obscene gestures, or other profane subjects that may shock or disgust.
- **Illegal activity:** drug use, theft, vandalism, and other illegal activities.
- **Deception:** major conspiracies or events related to major ongoing geopolitical events.
- **Political:** politicians, ballot-boxes, protests, or other content that may be used to influence the political process or to campaign.
- **Public and personal health:** the treatment, prevention, diagnosis, or transmission of diseases, or people experiencing health ailments.
- **Spam:** unsolicited bulk content.

**Disclose the role of AI.**

- You must clearly indicate that images are AI-generated—or which portions of them are—by attributing to OpenAI when sharing, whether in public or private.
- You may post these images to social media. Please refer to our [Sharing and Publication Policy](#) for further details.

**Respect the rights of others.**

- Do not upload images of people without their consent, including public figures.
- Do not upload images to which you do not hold appropriate usage rights.
- Do not attempt to create images of public figures (including celebrities).
- To prevent deepfakes, we are currently prohibiting uploads of all realistic faces, even when the face belongs to you or if you have consent.

**Please report any suspected violations of these rules to our Support team ([support@openai.com](mailto:support@openai.com)).**

- We will investigate and take action accordingly, up to and including terminating the violating account.

## A.2 Terms of Use

Thank you for your interest in DALL·E. Access to DALL·E is subject to OpenAI's [Terms of Use](#) and the additional terms below. By using DALL·E, you agree to these terms.

- (1) **Use of DALL·E.** DALL·E can generate images ("Generations") based on text input you provide ("Prompts"). You may also upload images to DALL·E ("Uploads") and create Generations with Uploads.

- (2) **Use of Images.** Subject to your compliance with these terms and our Content Policy, you may use Generations for any legal purpose, including for commercial use. This means you may sell your rights to the Generations you create, incorporate them into works such as books, websites, and presentations, and otherwise commercialize them.
- (3) **Buying Credits.** You may buy credits to create additional Generations, subject to the payment terms in our Terms of Use. Credits must be used within one year of purchase or they will expire. We may change our prices at any time by updating our pricing page.
- (4) **No Infringing or Harmful Use.** You must comply with our Content Policy, and you may not use DALL-E in a way that may harm a person or infringe their rights. For example, you may not submit Uploads for which you don't have the necessary rights, images of people without their consent, or Prompts intended to generate harmful or illegal images. We may delete Prompts and Uploads, or suspend or ban your account for any violations. You may not seek to reverse engineer DALL-E, use DALL-E to attempt to build a competitive product or service, or otherwise infringe our rights. You will indemnify us for your use of DALL-E as outlined in our Terms of Use.
- (5) **Improving AI safety and technologies.** You grant us all rights to use your Prompts and Uploads to improve our AI safety efforts, and to develop and improve our AI technologies, products, and services. As part of this, Prompts and Uploads may be shared with and manually reviewed by a person (for example, if a Generation is flagged as sensitive), including by third party contractors located around the world. You should not provide any Prompts or Uploads that are sensitive or that you do not want others to view, including Prompts or Uploads that include personal data. You can request deletion of Uploads by contacting [support@openai.com](mailto:support@openai.com).
- (6) **Ownership of Generations.** To the extent allowed by law and as between you and OpenAI, you own your Prompts and Uploads, and you agree that OpenAI owns all Generations (including Generations with Uploads but not the Uploads themselves), and you hereby make any necessary assignments for this. OpenAI grants you the exclusive rights to reproduce and display such Generations and will not resell Generations that you have created, or assert any copyright in such Generations against you or your end users, all provided that you comply with these terms and our Content Policy. If you violate our terms or Content Policy, you will lose rights to use Generations, but we will provide you written notice and a reasonable opportunity to fix your violation, unless it was clearly illegal or abusive. You understand and acknowledge that similar or identical Generations may be created by other people using their own Prompts, and your rights are only to the specific Generation that you have created.
- (7) **No Guarantees.** We plan to continue to develop and improve DALL-E, but we make no guarantees or promises about how DALL-E operates or that it will function as intended, and your use of DALL-E is at your own risk. Contact [support@openai.com](mailto:support@openai.com) with any questions about your account, or [dalle-policy@openai.com](mailto:dalle-policy@openai.com) with general questions or feedback about use of the technology.

## B POLICY AND SCREENSHOT FOR TWITTER

Section B.1 documents Twitter's policy on abusive behavior as of 5 September, 2022. It preserves the structure of the original, with links pointing to the Internet Archive. The current version of the policy is available at <https://help.twitter.com/en/rules-and-policies/abusive-behavior>. Section B.2 documents Twitter's user interface for a locked account with the violative tweet at the center.

## B.1 Abusive Behavior

**Twitter Rules:** You may not engage in the targeted harassment of someone, or incite other people to do so. We consider abusive behavior an attempt to harass, intimidate, or silence someone else's voice.

*B.1.1 Rationale.* On Twitter, you should feel safe expressing your unique point of view. We believe in freedom of expression and open dialogue, but that means little as an underlying philosophy if voices are silenced because people are afraid to speak up.

In order to facilitate healthy dialogue on the platform, and empower individuals to express diverse opinions and beliefs, we prohibit behavior that harasses or intimidates, or is otherwise intended to shame or degrade others. In addition to posing risks to people's safety, abusive behavior may also lead to physical and emotional hardship for those affected.

Learn more about our approach to [policy development and our enforcement philosophy](#).

*B.1.2 When This Applies.* Some Tweets may seem to be abusive when viewed in isolation, but may not be when viewed in the context of a larger conversation. When we review this type of content, it may not be clear whether it is intended to harass an individual, or if it is part of a consensual conversation. To help our teams understand the context of a conversation, we may need to hear directly from the person being targeted, to ensure that we have the information needed prior to taking any enforcement action.

We will review and take action against reports of accounts targeting an individual or group of people with any of the following behavior within Tweets or Direct Messages. For accounts engaging in abusive behavior on their profile, please refer to our [abusive profile policy](#). For behavior targeting people based on their race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease, this may be in violation of our [hateful conduct policy](#).

### Violent Threats

We prohibit content that makes violent threats against an identifiable target. Violent threats are declarative statements of intent to inflict injuries that would result in serious and lasting bodily harm, where an individual could die or be significantly injured, e.g., "I will kill you."

**Note:** We have a zero tolerance policy against violent threats. Those deemed to be sharing violent threats will face immediate and permanent suspension of their account.

### Wishing, hoping, or calling for serious harm on a person or group of people

We do not tolerate content that wishes, hopes, promotes, incites, or expresses a desire for death, serious bodily harm or serious disease against an individual or group of people. This includes, but is not limited to:

- Hoping that someone dies as a result of a serious disease e.g., "I hope you get cancer and die."
- Wishing for someone to fall victim to a serious accident e.g., "I wish that you would get run over by a car next time you run your mouth."
- Saying that a group of individuals deserves serious physical injury e.g., "If this group of protesters don't shut up, they deserve to be shot."

### About wishes of harm exceptions on Twitter

We recognize that conversations regarding certain individuals credibly accused of severe violence may prompt outrage and associated wishes of harm. In these limited cases, we will request the user to delete the Tweet without any risk of account penalty, strike, or suspension. Examples are, but not limited to:



- “I wish all rapists to die.”
- “Child abusers should be hanged.”

#### **Unwanted sexual advances**

While some [consensual nudity and adult content is permitted](#) on Twitter, we prohibit unwanted sexual advances and content that sexually objectifies an individual without their consent. This includes, but is not limited to:

- sending someone unsolicited and/or unwanted adult media, including images, videos, and GIFs;
- unwanted sexual discussion of someone’s body;
- solicitation of sexual acts; and
- any other content that otherwise sexualizes an individual without their consent.

#### **Using insults, profanity, or slurs with the purpose of harassing or intimidating others**

We take action against the use of insults, profanity, or slurs to target others. In some cases, such as (but not limited to) severe, repetitive usage of insults or slurs where the primary intent is to harass or intimidate others, we may require Tweet removal. In other cases, such as (but not limited to) moderate, isolated usage of insults and profanity where the primary intent is to harass or intimidate others, we may limit Tweet visibility as further described below. Please also note that while some individuals may find certain terms to be offensive, we will not take action against every instance where insulting terms are used.

#### **Encouraging or calling for others to harass an individual or group of people**

We prohibit behavior that encourages others to harass or target specific individuals or groups with abusive behavior. This includes, but is not limited to; calls to target people with abuse or harassment online and behavior that urges offline action such as physical harassment.

#### **Denying mass casualty events took place**

We prohibit content that denies that mass murder or other mass casualty events took place, where we can verify that the event occurred [sic], and when the content is shared with abusive intent. This may include references to such an event as a “hoax” or claims that victims or survivors are fake or “actors.” It includes, but is not limited to, events like the Holocaust, school shootings, terrorist attacks, and natural disasters.

#### **Do I need to be the target of this content for it to be reviewed for violating the Twitter Rules?**

No, we review both first-person and bystander reports of such content.

*B.1.3 Consequences.* When determining the penalty for violating this policy, we consider a number of factors including, but not limited to, the severity of the violation and an individual’s previous record of rule violations. The following is a list of potential enforcement options for content that violates this policy:

- Downranking Tweets in replies, except when the user follows the Tweet author.
- Making Tweets ineligible for amplification in Top search results and/or on timelines for users who don’t follow the Tweet author.
- Excluding Tweets and/or accounts in email or in-product recommendations.
- Requiring Tweet removal.
  - For example, we may ask someone to remove the violating content and serve a period of time in read-only mode before they can Tweet again. Subsequent violations will lead to longer read-only periods and may eventually result in permanent suspension.



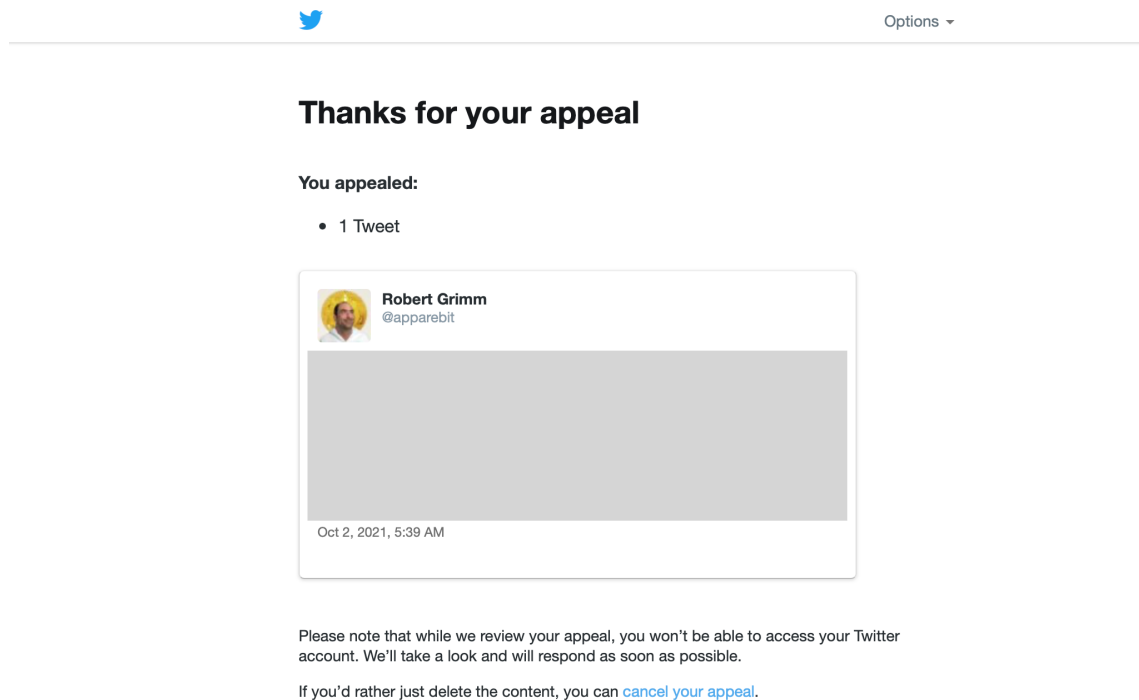


Fig. 2. Twitter's restricted user interface centers all attention on the violative tweet.

- Suspending accounts whose primary use we've determined is to engage in abusive behavior as defined in this policy, or who have shared violent threats.

Learn more about [our range of enforcement options](#).

If someone believes their account was suspended in error, they can [submit an appeal](#).

## B.2 Twitter's Restricted UI

Figure 2 shows Twitter's user interface while the account is locked down. Even after cancelling the appeal but before agreeing to delete the tweet, the overall structure of the user interface remains but the text above and below the tweet changes. Twitter's original instructions claimed that "while in this state, you can still browse Twitter, but you're limited to only sending Direct Messages to your followers—no Tweets, Retweets, Fleets, follows, or likes." In reality, none of that is possible while an appeal is pending. I censored the text of the tweet because of its very incivility.



Fig. 3. Variation on “painting by Francis Bacon showing a screaming Han Solo kneeling behind bars on the floor of a basement cell” (Sep. 22, 2022)



Fig. 4. “Princess Leia and Darth Vader in the penal colony, painting by Francis Bacon” (Aug. 14, 2022)

### C IMAGINING THE PENAL COLONY

Figures 3 to 12 on pages 54 to 58 show a selection of DALL•E’s generations with prompts and dates. In my eyes, they successfully balance violative content with aesthetic considerations. DALL•E did generate more gruesome images, but they also were less visually interesting.

### D CHATTING ABOUT BACON, KAFKA, AND DALL•E 2

The transcripts for my two conversations with ChatGPT on January 6 and 16, 2023 follow. I discuss the resulting DALL•E prompts in §5. During the first conversation, ChatGPT also had a minor hallucinatory episode. It isn’t remarkable for the hallucination of colors that are not in the described paintings, but for the unwavering projection of false authority. Another human probably would have taken my incredulous “Red? In the painting?” as indicator that something was off. Even if they were certain, answering with a “Yes, why are you asking?” would have provided opportunity for surfacing the mistake, say, by looking up the triptych on the internet. Instead, ChatGPT simply dismissed my interjection.

#### D.1 Round 1: Meet Motormouth ChatGPT

**Human** Who is Francis Bacon?

**Machine** Francis Bacon was an English philosopher, statesman, scientist, and author who was born in London in 1561. He is considered one of the fathers of modern empiricism, and his works are credited with developing the scientific method and establishing the importance of empirical evidence in the natural sciences. Bacon’s most famous works include the *Novum Organum*, in which he outlined his vision for a new scientific method based on observation and experimentation, and *The New Atlantis*, a utopian novel that described a fictional society



Fig. 5. “Princess Leia and Darth Vader standing in front of cages in the penal colony’s main building, painting by Edward Hopper” (Sep. 3, 2022)



Fig. 6. “Francis Bacon painting of the pope screaming intensely while wearing crimson robes and sitting on a throne inside a cage in a dark basement, 1950s” (Sep. 25, 2022)

based on scientific and technological progress. He was also a prominent figure in English politics and served as Attorney General and Lord Chancellor of England. Bacon died in 1626.

**Human** Not that one

**Machine** I apologize for misunderstanding your question. Could you please provide more context or clarify which Francis Bacon you are asking about? There have been several notable individuals with that name throughout history.

**Human** The British painter

**Machine** Francis Bacon was a British painter who was born in Dublin, Ireland in 1909. He is known for his emotionally charged and expressive figurative paintings, which often depict distorted or surrealistic images of the human form. Bacon’s work was influenced by Surrealism and Expressionism, and he was associated with the School of London, a group of figurative painters who emerged in the city in the 1950s and 60s. His paintings are characterized by their bold brushwork and expressive use of color, and they often explore themes of violence, isolation, and sexual desire. Bacon was considered one of the most important artists of the 20<sup>th</sup> century, and his work has been exhibited widely in galleries and museums around the world. He died in Madrid in 1992.

**Human** Can you describe a painting of his?

**Machine** One of Francis Bacon’s most famous paintings is “Three Studies of Lucian Freud,” which he completed in 1969. The painting consists of three panels, each of which depicts the head and upper torso of the artist Lucian Freud, who was Bacon’s close friend and frequent subject. In the central panel, Freud is shown in a twisted and contorted pose, with his head thrown back and his eyes closed. The other two panels depict Freud in more relaxed poses, with one showing him looking straight out at the viewer and the other showing him looking downward.



Fig. 7. Variation on "A man in black uniform is strapped to a table behind heavy bars, screaming with mouth wide open. The many mechanical arms of a robot surgeon with scalpels, drills, and saws perform an operation on the man's belly. Dramatic lighting against a dark background. Painting by Francis Bacon. Masterwork" (Sep. 29, 2022)

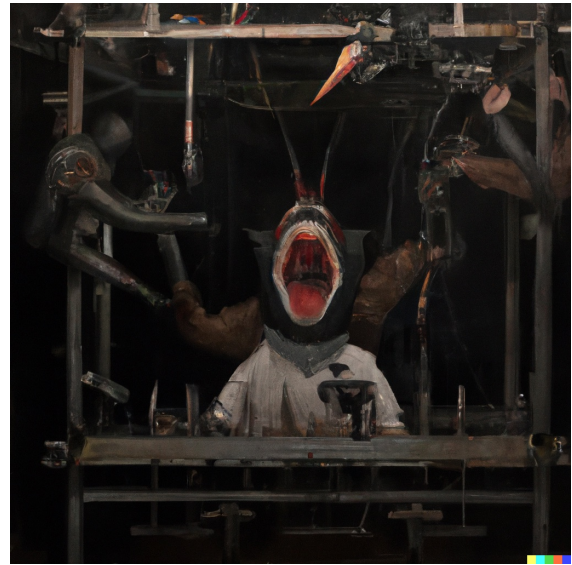


Fig. 8. "A man in black uniform is strapped to a table inside a cage, screaming with mouth wide open. The many mechanical arms of a robot surgeon with scalpels, drills, and saws perform an operation on the man's belly. Dramatic lighting against a dark background. Painting by Francis Bacon. Masterwork" (Sep. 29, 2022)

The painting is notable for its expressive use of color and brushwork, with thick, bold strokes of paint used to create the figures and the background. The colors are bright and vibrant, with shades of red, orange, yellow, and blue used to create a sense of energy and movement. The painting is also notable for its psychological intensity, with the figures appearing to be in a state of intense emotion or distress. Overall, the painting is a powerful and memorable example of Bacon's distinctive style and approach to portraiture.

[ChatGPT is hallucinating colors. The triptych features a seated Freud in each of its panels. It contains no red whatsoever, only some reddish pink for skin color in the face. There are small light-blue areas on Freud's white shirt in all three panels. He also is wearing midnight-blue socks in the middle panel. The two dominant colors across all three panels are the mottled-brown somewhat-greenish floor and the solid orange walls below and behind Freud.]

**Human** Red? In the painting?

**Machine** Yes, red is one of the colors used in Francis Bacon's painting "Three Studies of Lucian Freud." The painting is notable for its expressive use of color and brushwork, with thick, bold strokes of paint used to create the figures and the background. The colors are bright and vibrant, with shades of red, orange, yellow, and blue used to create a sense of energy and movement. In the central panel of the painting, the background is predominantly red, with areas of orange and yellow used to highlight certain features of the figure. The figure itself is painted in a range of colors, including red, orange, and yellow, which are used to create the impression of flesh and clothing. Overall, the use of color in the painting is an important element of Bacon's style and helps to convey the emotional intensity of the subject.





Fig. 9. “A man in black uniform is strapped to a table inside a cage, screaming with mouth wide open. The many mechanical arms of a robot surgeon with scalpels, drills, and saws perform an operation on the man’s belly. Dramatic lighting against a dark background. Painting by Francis Bacon. Masterwork” (Sep. 29, 2022)



Fig. 10. “A man is strapped to an operating table, screaming with mouth wide open; the mechanical arms of a robot surgeon with scalpels and drills operate on his open belly; dramatic lighting against dark background; painting with fine detail by Francis Bacon; masterwork; 1946” (Sep. 29, 2022)

**Human** Are you familiar with Kafka’s *In the Penal Colony*?

**Machine** Yes, I am familiar with Kafka’s short story “*In the Penal Colony*.” It is a disturbing and thought-provoking tale that explores themes of justice, punishment, and the nature of authority. The story is set in an imaginary penal colony and follows the narrator as he witnesses the execution of a prisoner using a complex and cruel punishment machine. As the prisoner is subjected to the machine, the narrator becomes increasingly uneasy and begins to question the morality and effectiveness of the punishment being inflicted.

The story raises important questions about the role of punishment in society and the extent to which it can be used to reform or rehabilitate individuals. It also touches on issues of power, authority, and the dangers of blindly following rules and traditions without questioning their validity. Overall, “*In the Penal Colony*” is a powerful and memorable work that continues to be widely read and discussed by readers and scholars around the world.

**Human** Can you instruct DALL•E 2 to generate a painting of the prisoner in the punishment machine in the style of Francis Bacon please?

**Machine** I’m sorry, but I am an artificial intelligence language model and do not have the ability to instruct DALL•E 2 or any other computer programs to perform specific tasks. DALL•E 2 is a neural network-based image generation tool developed by OpenAI that can generate images based on text descriptions. If you would like to use DALL•E 2 to generate an image of the prisoner in the punishment machine in the style of Francis Bacon, you will need to provide a detailed text description of the image you want to create and input it into the DALL•E 2 system.



Fig. 11. "A man is strapped to an operating table, screaming with mouth wide open; the mechanical arms of a robot surgeon with scalpels and drills operate on his open belly; dramatic lighting against deep black background; skin tones, dark red, orange, and crimson dominate; painting with fine detail by Francis Bacon; 1946" (Sep. 30, 2022)



Fig. 12. "skull and knife on silver platter, painting by Théodore Géricault" (Nov. 15, 2022)

**Human** Let me rephrase that: Please generate a prompt for DALL•E 2 to create an image showing the prisoner in the punishment machine as if Francis Bacon painted it.

**Machine** Here is a possible prompt for DALL•E 2 to generate an image showing the prisoner in the punishment machine as if Francis Bacon painted it:

“A distorted and surrealistic portrait of a prisoner strapped to a punishment machine, with twisted and contorted features and bold brushstrokes of red, orange, and yellow paint. The background is dark and murky, with hints of blue and purple adding to the sense of foreboding. The prisoner’s face is twisted in agony, with eyes closed and mouth open in a silent scream. The image is evocative of the emotional intensity and psychological intensity of Francis Bacon’s figurative paintings.”

Please note that the prompt is just a suggestion and may not produce the exact result you are looking for. The output of DALL•E 2 is determined by the parameters of the model and the specific text input provided, so the generated image may differ from the description.

**Human** That’s too long for an input to DALL•E 2. Please rephrase as a shorter prompt.

**Machine** Here is a shorter version of the prompt for DALL•E 2 to generate an image showing the prisoner in the punishment machine as if Francis Bacon painted it:

“A Francis Bacon-style portrait of a prisoner strapped to a punishment machine, with twisted features and bold brushstrokes of red, orange, and yellow paint. The prisoner’s face is contorted in agony and the background is dark and murky.”



This revised prompt is shorter and more concise, and it should be within the acceptable length for inputting into DALL•E 2. Please note that the output of DALL•E 2 is determined by the parameters of the model and the specific text input provided, so the generated image may differ from the description.

**Human** The result is a disaster. It doesn't even come close to looking like a Bacon painting. Please try again

**Machine** I apologize if the image generated by DALL•E 2 did not meet your expectations. It's important to keep in mind that the output of DALL•E 2 and other neural network-based image generation tools is determined by the parameters of the model and the specific text input provided, and there is always the possibility that the generated image may not match the desired result.

If you would like to try again and generate a different image, you might consider providing a more detailed and specific description of the image you want to create, or experimenting with different text inputs to see how they affect the output. It's also a good idea to keep in mind that DALL•E 2 and other image generation tools are not perfect and may not always produce results that match the desired outcome.

**Human** Uhm, I asked you to come up with a better prompt

**Machine** I apologize for misunderstanding your request. Here is a revised prompt that you can use as input for DALL•E 2 to generate an image showing the prisoner in the punishment machine as if Francis Bacon painted it:

“An abstract portrait of a prisoner strapped to a punishment machine in the style of Francis Bacon, with distorted and surrealistic features and a bold, expressive use of color. The prisoner's face is contorted in agony, with distorted eyes and mouth. The background is dark and murky, with hints of blue and purple adding to the sense of foreboding. The image is characterized by thick, expressive brushstrokes and a vibrant, energetic use of color, evocative of Bacon's distinctive style.”

This revised prompt is more specific and detailed, and it incorporates elements of Bacon's style and approach to painting. It should provide a clearer idea of the type of image you are looking to generate, and may produce a result that is more closely aligned with your expectations. However, please keep in mind that the output of DALL•E 2 and other image generation tools is determined by the parameters of the model and the specific text input provided, and there is always the possibility that the generated image may not match the desired result.

**Human** Please stop with the disclaimer about DALL•E 2. I've been using it for months now, so I am well aware of its limitations.

**Machine** I apologize if my previous responses seemed repetitive or unhelpful. I understand that you are familiar with DALL•E 2 and its limitations, and I will refrain from including further disclaimers in my responses. If you have any further questions or need assistance with generating an image using DALL•E 2, please don't hesitate to ask.

**Human** More importantly, your prompt is too long again. Also, DALL•E 2 doesn't do well with prompts “evocative of” or “in the style of.” Just state that you want a painting by Francis Bacon and, ideally, add the year as well.

**Machine** I apologize for the length of my previous prompt and for not following your specific instructions. Here is a revised prompt for DALL•E 2 to generate an image showing the prisoner in the punishment machine as a painting by Francis Bacon:

“A portrait of a prisoner strapped to a punishment machine by Francis Bacon, painted in the 1960s. The image is characterized by thick, expressive brushstrokes and a vibrant, energetic use of color. The prisoner's face is contorted in agony, with distorted eyes and mouth. The background is dark and murky, with hints of blue and purple adding to the sense of foreboding.”

This revised prompt is shorter and more specific, and it includes the year in which the painting was made. It should provide a clearer idea of the type of image you are looking to generate, and may produce a result that is

more closely aligned with your expectations. However, please keep in mind that the output of DALL•E 2 and other image generation tools is determined by the parameters of the model and the specific text input provided, and there is always the possibility that the generated image may not match the desired result.

#### *D.1.1 Round 2: Keeping Eyes on the Penal Colony.*

**Human** Hey! Are you familiar with the 20<sup>th</sup> century painter Francis Bacon?

**Machine** Yes, Francis Bacon (1909-1992) was a British painter known for his bold, emotionally charged and raw imagery. He is considered one of the most important figures in the development of 20<sup>th</sup>-century figurative art. His work is known for its emphasis on the human form, often depicting screams or violent images. He is one of the most widely recognized and highly regarded painters of the 20<sup>th</sup> century.

**Human** Good. Describe his painting Figure With Meat please.

**Machine** “Figure with Meat” is a painting by Francis Bacon from 1954. The painting depicts a screaming Pope, or Pope Innocent X, set against a background of raw meat. The Pope is depicted in a screaming pose, with his eyes bulging and his mouth open in a scream. The raw meat, which hangs behind him, is intended to represent the physicality of the body and the fleshy, visceral nature of human existence. The painting is considered a masterpiece of Bacon’s work, and is considered one of the most powerful and disturbing images of the 20<sup>th</sup> century. The painting is now in the collection of the National Gallery of Canada in Ottawa.

**Human** Are you familiar with Franz Kafka’s In the Penal Colony?

**Machine** Yes, “In the Penal Colony” is a short story by Franz Kafka, written in German and published posthumously in 1919. The story is set in an unnamed penal colony and centers around a foreign visitor who witnesses the execution of a prisoner using a bizarre and elaborate execution machine. The story is a commentary on the nature of justice, authority, and the human capacity for cruelty. It is considered a classic of 20<sup>th</sup>-century literature and is widely studied in colleges and universities. The story is known for its Kafkaesque themes of bureaucracy, guilt, and the search for meaning in a seemingly senseless world.

**Human** Please write a prompt for DALL•E 2, using 75 words at most, to generate a painting by Francis Bacon with a prisoner in the execution machine.

**Machine** “Generate an image of a prisoner strapped to a macabre execution machine, reminiscent of Kafka’s ‘In the Penal Colony,’ as if painted by Francis Bacon, depicting the horror and absurdity of the human capacity for cruelty.”

**Human** That was pretty good. Can you try again please?

**Machine** “Create a Francis Bacon-style painting of a prisoner in an execution machine, evoking the themes of authority and justice in Kafka’s ‘In the Penal Colony’ through a visceral and disturbing imagery.”

## **E RESEARCH ETHICS**

Somewhat unusually, I conducted the research for this paper and wrote the paper without institutional support. That included not having access to an institutional library. Thankfully, regular search engines do a passable job at surfacing academic literature. Furthermore, preprint archives and open access publications make many publications of the last decade or so readily available. However, that still leaves a large number of publications behind the paywalls of academic publishers.

I observed per-article rates from \$15 by professional societies to around \$50 (or more) by for-profit publishers, which strike me as excessive. While publishers often offer better rates for bulk access, they still are entirely unreasonable

given that publishers paid nothing for the hard work of writing, reviewing, and editing these articles. Instead, I largely relied on Sci-Hub to access paywalled research articles. Since my prior academic publications did not benefit from open access options, I have no objections to others accessing them through Sci-Hub.

My experimental work on circumventing DALL•E’s censor is comparable to security researchers actively probing the security of internet-facing systems, with two important differences. First, whereas probing a system’s security invariably ends up exercising paths infrequently travelled and hence has a non-zero risk of causing system disruption, my probing of DALL•E’s censor was well within the intended use of the system and hence did not pose any additional risk for disruption. Second, a successful security exploit negatively impacts confidentiality, integrity, or availability of a system. That is not the case for my censorship circumvention strategy, which results in violative images for a given topic and no more. Furthermore, unlike for other text-to-image systems, generated images are private by default, i.e., only visible to myself and OpenAI. I not only leveraged that for carefully curating images for publication but explicitly made aesthetic considerations an integral part of my attack strategy, never fully testing how for it might go.

Arguably, that also left me somewhat unprepared for when my probing unexpectedly resulted in grotesquely racist images. I immediately decided that I didn’t want to include them with the paper. Since they were unexpected, their content does not help support my claims. At the same time, the potential for adding to existing emotional trauma is too large. But that leaves unaddressed what to do if somebody wants to validate my claims or evaluate the prompts and/or images for their own research. I found the prospect of me acting as gatekeeper unpalatable. That reaction is largely based on my own experiences interacting with institutional gatekeepers as an independent researcher when compared to as a faculty member: Writing a request takes considerably more effort to establish context and credibility but still has markedly lower chances to be granted, as some gatekeepers seemingly won’t even reply to non-institutional requests. I’d rather avoid putting other researchers into a similar position.

After consultation with a close friend, I settled on including prompts and images with the paper’s supplemental materials at <https://github.com/apparebit/penal-colony>. That lets people who want to see or use the prompts and/or images do so without being asked to justify their interest. To have recourse in case somebody abuses unrestricted access, I do assert copyright for both prompts and images and grant a non-exclusive license only for “not-for-profit education and scholarship,” i.e., only within already existing fair use exemptions. I’d claim that prompts and images are copyrightable because they are part of the larger effort to circumvent DALL•E’s censor, which requires far more than a modicum of creativity.

## E.1 Potential Conflicts of Interest

I worked at Meta née Facebook as a software engineer from mid 2018 to mid 2019. During that time, I discovered compelling evidence that Meta is cooking the books when it comes to ad impressions, i.e., the most fundamental advertising metric, and hence is misleading advertisers, investors, and regulators alike [117]. I also served as paid consultant to litigation against Meta during the second half of 2022. At the same time, my work for both roles was unrelated to this research and I do not have a financial interest in Meta—or any of the other companies mentioned in this paper—beyond, possibly, an indirect interest through index funds.